Open Science at Web-Scale: Optimising Participation and Predictive Potential

Consultative Report

Document details

Author:	Dr Liz Lyon, UKOLN/ Digital Curation Centre, University of Bath
Date:	6th November 2009
Version:	V1.0
Document Name:	open-science-report-6nov09-final-sentojisc.doc
Notes:	Revised after comments from JISC and reviewers

Acknowledgement to contributors

The author would like to thank the various people, who contributed to the report by completing an interview, making a presentation, or commenting on previous versions. The author takes responsibility for interpreting the answers and for any change of emphasis that comes with collating the viewpoints of the various contributors.

Acknowledgement to funders

This work was funded by the JISC as part of the Information Environment programme.

UKOLN is funded by the MLA: The Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils, as well as by project funding from the JISC, the Research Councils and the European Union. UKOLN also receives support from the University of Bath where it is based.

Dedicated to
Rosalind Franklin Ph.D. (1920-1958)

1 Executive Summary

This Report has attempted to draw together and synthesise evidence and opinion associated with data-intensive open science from a wide range of sources. The potential impact of data-intensive open science on research practice and research outcomes, is both substantive and far-reaching. There are implications for funding organisations, for research and information communities and for higher education institutions.

The original specification for the work was highly selective in its choice of areas to study, and this Report addresses only three of these areas in any depth:

- *open science including open notebook science :* making methodologies, data and results available on the Internet, through transparent working practices
- *citizen science including volunteer computing :* where volunteers who may not have scientific training, perform or manage research-related tasks such as observation, measurement or computation
- *predictive science :* data-driven science which enables the forecasting, anticipation or prediction of specific outcomes.

Synthetic science (research which combines science and engineering methods to design and build novel biological entities), and Immersive science (used to describe research involving virtual and simulated worlds), are referenced, but require more detailed examination. Fuller definitions of the terms and areas examined in this study have been provided in Section 3. In addition, the Report addresses data informatics and the supporting role of libraries for these particular aspects of open science.

The work was undertaken through a mix of desk research, including analysis from the peer-reviewed literature, presentations, selective blogs, wiki content, social network discussion, and by consultation with a small group of leading thinkers and researchers. The Report was also informed by presentations and talks given by the author during 2009.

The Report is positioned as a Consultative document, which it is hoped will stimulate and contribute to community discussion in the UK, but also fuel the open science debate on the global stage. Whilst many questions have been asked here, they will require fuller articulation and investigation in other fora. The economic implications will require detailed analysis and the societal benefits should be reviewed and evaluated. The consultative questions are clearly indicated in boxes in the text and are reproduced in full in the Executive Summary.

Consultation Challenge 1 : Scale, Complexity and Predictive Potential

Data-intensive science powered by contemporary computational hardware, software and research techniques, enables scientists to perform experiments and calculations at different orders of magnitude of scale and volume: research that was completed in a year can now be repeated in a weekend. Sustained growth in data modelling, complex simulations and visualisations, facilitate interpretation and analysis by humans and machines, leading to the development of predictive science scenarios in a wider range of disciplines. Examples of data intensive science at these extremes of scale, which enable forecasting and predictive assertions, have been described.

Assessments of the accuracy and robustness of predictions are linked to uncertainty quantification, the accuracy of the underlying model, and the integrity of the data. Key questions address community awareness and understanding of the potential implications and impact of (open) data-intensive science at new extremes of scale and complexity, and the service requirements for associated data curation and preservation.

What is the level of awareness and understanding in the wider community of the prospects and societal implications of predictive science?

How are the methodologies and tools for data quality, validation and verification, which underpin robust and trustworthy large-scale models and simulations, implemented in different disciplines? Are appropriate data quality standards in place?

How are the necessary mathematical skills available to science teams, particularly in domains such as biology?

How can services like the Digital Curation Centre, best support the effective curation and long-term preservation of complex and dynamic data models, simulations and visualisations?

Consultation Challenge 2 : Continuum of Openness

Open science has been presented in this Report as a continuum, which is helpful in positioning the range of behaviours and practices observed in different disciplines and contexts. The twin aspects of openness (access and participation), have been separated to facilitate scoping the full potential of the open science vision and a listing of the perceived values and benefits of open science is given. Available evidence suggests that transparent data sharing and data re-use are far from commonplace and some of the reasons for this are examined. Peer production approaches to data curation are in their infancy but offer considerable promise as scaleable models which could be migrated to other disciplines. The more radical open notebook science methodologies are currently on the “fringe” and it is not clear whether uptake and adoption will grow in other disciplines and contexts. The challenge of “openness” across its range of interpretations, demands that we address the awareness and understanding of fundamental open science concepts, supplemented by probing exploration of practitioner experience.

What are the views of the community on open science principles, acknowledging that “openness” is a continuum or sliding scale with different groups, services, information and data, positioned at different points?

What are the views of the community on the perceived value and benefits of open science methodologies? How can these benefits be demonstrated and evaluated?

Should research funding bodies be pro-actively supporting open science principles and practice? What are the policy implications? What infrastructure is required?

How aware are the majority of scientists of the range of social Web tools available to support open science? How are the tools used in different disciplines? What are the perceived advantages and disadvantages of using collaborative tools? How can social tools add value to research? What are the cost-benefits of using these types of tools?

What are the implications of open science communication channels e.g. blogs, on scholarly publishing models? What are the views of publishers and learned societies?

How can the peer production model for data curation, be applied and adopted in other disciplines?

What are the community views on Open Notebook Science? Should these radical methods be migrated across to other disciplines and if so, which other disciplines would benefit? What key ONS development and enhancement issues need to be addressed?

Consultation Challenge 3 : Citizen Science

21st Century team science has been empowered by the proliferation of social Web tools enabling globally distributed groups to work together, but we can also envisage team science embracing interested amateurs and citizens, as well as research professionals. Some established and compelling exemplars of citizen science are given, but it is noted that this model may be more suited to certain domains and types of research. However, the growth of mobile phone use in citizen journalism, for public census work and participative surveys and the the development of sensor-rich mobile devices, suggest that there is great potential for more participatory methodologies to benefit scientific research, though some significant privacy and legislative issues remain unanswered.

The influence of computer gaming approaches to motivate participants in volunteer computing initiatives is described, and the development of citizen science Web services, system architectures and the design of appropriate interfaces, is briefly explored. We need to learn much more about how the public interact with these services to maximise the value and benefit from such investment. The basic questions probing citizen science, raise significant philosophical and pragmatic issues for professional scientists, research funding bodies, higher education institutions and the wider community.

What are scientist and funder attitudes towards citizen science? What are the societal implications? What role should research funding bodies play?

What are the short, medium and long term strategic and policy implications on science practice and outcomes, of a more openly participative research approach which may pro-actively include the public?

What are the financial implications, both in terms of direct and indirect costs, investment in infrastructure and associated benefits? What are the risks? What is the impact on research quality (data, models, outcomes)?

Which disciplines and areas of research are most suited to citizen science methodologies? How should the collaboration market model be applied to research?

How will open and participative science initiatives impact on research practice in HE institutions? How should professional scientists, volunteers, amateurs and citizen scientists (and all flavours in between), work together in a socially optimal manner where there is mutual benefit? What can scientists learn from citizen journalism?

What are the technical requirements for designing effective citizen science Web services and systems? What can we learn from current successful exemplars?

Consultation Challenge 4 : Credentials, Incentives and Rewards

The potential impact of these changing practices on established business models for science and scholarly communications is raised: new notions of reputation and trust are developing which challenge established norms. There is brief discussion of the current journal publishing model with associated citation metrics for UK research assessment, which does not reward data sharing, social Web contributions or peer production approaches to data curation. Some novel proposals which seek to include such parameters in research assessment metrics are presented. The implications on research funder policies, future science investment planning and scholarly communication business models are not fully understood, but it is clear that the lack of incentives for data sharing and participatory methodologies, are a barrier to the wider adoption of the open science agenda. The consultative questions explore incentivising data sharing and re-use, and strategies for enabling more open participation, in the context of open science and scholarly communications.

Should open science practices be formally recognised and rewarded as intrinsic elements of scholarly communications? How can this be best achieved?

What are the views of the research community on appropriate incentives and reward structures for data sharing, data re-use and wider participation?

What are the views of the research funding bodies? Should these types of contribution and associated metrics, be included in future research assessment frameworks? How should they be assessed? How is the proposed Scholar Factor perceived? How should such metrics supplement journal citation metrics?

What are the views of scholarly publishers and learned societies? How do these contribution channels affect scholarly communication business models?

Consultation Challenge 5 : Institutional Readiness and Response

The open science agenda, with the data-intensive science at extremes of scale described in this Report, has significant implications for higher education institutions at policy, planning and operational levels. This Report raises some preliminary points and an Open Science Institutional Readiness Checklist is given as a brief aide memoire for institutions. It is hoped that by asking basic questions which explore institutional awareness, policy, planning and research practice, the community will begin to explore these substantive issues in more depth.

How aware are institutional senior management teams of the strategic implications of this potentially transformational agenda? How can research funding organisations, the JISC and other research support bodies help to raise awareness amongst institutional leaders? Who will lead and co-ordinate this work? What can be leveraged by partnerships on a global scale?

What are the implications for investment in research infrastructure? What can private sector organisations including ICT companies, contribute? What partnership opportunities arise?

How will academic structures evolve to support data-intensive science at extremes of scale? What institutional policy implications arise from open science practice? How are open scholarly communications channels such as research blogs supported in HEIs? Where are institutions positioned on open data-sharing? What are the IPR issues? What are the policy implications for institutions, of co-working with non-professionals i.e. volunteers and interested amateurs? What are the societal benefits?

What guidance is provided for research staff? How are open science issues and practices, addressed in staff induction and professional development courses? How can advocacy materials for institutions (e.g. a Team Science Toolkit), help to provide guidance and support for planning, policy development and good working practices?

Consultation Challenge 6 : Data Informatics Capacity and Capability

Particular attention has been paid to the provision of data informatics capacity and capability and the role of the Library in this context. The Report asserts that Libraries are well-placed to support research data management but that new skills and roles will need to be embraced by the professional LIS community. Modifications to LIS courses will be required and there are

similar training implications for new-entrant researchers and postgraduates, to equip them with the skills and methodologies required for data-intensive science. The UK Digital Curation Centre is a key resource, although the increasing demands on this relatively modest service are challenging. The consultative questions explore the embedding of skills required for open data-intensive science, the role of the Library and Information Services and implications for postgraduate training and LIS curriculum development.

What is the research community view on the current provision of data informatics skills for postgraduates and research staff? If current curricula and training are not meeting needs, how can the position be improved? Should basic data informatics training be a core element of courses? Who should provide this training? What are the costs?

How can research funding agencies best support data informatics skills development?

What is the community perspective on the roles that Libraries and Information Services could play in supporting open data-intensive science? How can academic and research libraries be empowered to engage and participate in team science initiatives?

What is the role of SCONUL, RLUK, CILIP and other professional LIS organisations?

How should Library and Information Science schools address the provision of data curation and data informatics expertise within their courses and programmes?

Finally, it is intended that Recommendations for further work will arise from the subsequent community and stakeholder discussion.

2 Introduction

Recent dramatic changes in the participative character and social focus of the Web, together with the trend towards increasingly data-driven and "in silico" research, augmented by more open methodologies, have begun to influence scientific practice, research methodologies and ways of working. These trends have the potential to radically transform science, but also have significant practical implications for individuals, institutions and stakeholder organisations in the JISC arena. This study aims to describe and evaluate these changes, and in particular, to identify and articulate associated issues and challenges for further discussion by the wider community.

2.1 Terms of Reference and Objectives

As stated in the original brief, the consultancy objectives were:

- *To describe current practice in five selected emerging areas of science, chosen for their heavy reliance on data, information systems, distributed networks and computational resource, but also for their transformational potential and ability to impact on JISC work. The selected areas will include:*
    - *open science including open notebook science*
    - *citizen science including volunteer computing*
    - *predictive science*
    - *synthetic science*
    - *immersive science.*
- *To explore current limitations and barriers in each exemplar area, to highlight the inherent risks and to try to assess the potential for the future. Policy issues, technical challenges and socio-legal aspects together with relevance to scholarly communications, will be covered.*
- *To relate this changing science practice to stakeholder organisations and associated JISC activity. Stakeholder organisations would include libraries and information services, data centres and institutions, as well as the JISC and other research funders and policy makers. Some consideration would be given to the types of active community in these areas and whether and how JISC might work with them or implement similar models.*
- *To make appropriate Recommendations for further work.*

2.2 Audience

The primary audiences for the report are:

- the JISC Executive and Innovation Team
- the relevant JISC Committees
- the wider community.

The report will also be made available from the JISC and UKOLN/DCC Web sites.

2.3 Methodology

The work has been undertaken through a mix of desk research, including analysis from the peer-reviewed literature, presentations, selective blogs, wiki content, social network discussion, and by consultation with a small group of leading thinkers and researchers (see Appendix). This Report provides a synthesis of information and opinion gathered throughout the study, with additional analysis and commentary. The Report has also been informed by various presentations and talks given by the author during 2009.

2.4 Positioning and Scope

This Report has been positioned as a consultative document, which seeks to raise a number of issues and challenges for community comment and debate. Many of the issues raised are substantive items, having potentially far-reaching implications for funding organisations, for research and information communities and for higher education institutions. It is anticipated that selected aspects will require further investigation and exploration. The aim here, is simply to surface the issues and the consultative questions are clearly indicated in boxes in the text. It is intended that Recommendations for further work will arise from the subsequent discussion.

The original specification for this Report was highly selective in its choice of areas to study and following a long gestational period, this Report addresses only the first three of these substantive areas in any depth; the last two areas are referenced, but require more detailed examination. However in addition, the Report does address data informatics and libraries.

The body of the Report is arranged in nine sections: Definitions, Context, Scale Complexity and Predictive Potential, Continuum of Openness, Citizen Science, Credentials, Incentives and Rewards, Institutional Readiness and Response, Data Informatics Capacity and Capability, and Conclusions.

3 Definitions

For clarity, it will be helpful to present definitions of the terms and areas examined in this study.

Open Science: In his Peanutbutter blog, Frank Gibson wrote[1]:

"*Open Science" encompasses the ideals of transparent working practices across all of the life-science domains, to share and further scientific knowledge. It can also be thought of to include the complete and persistent access to the original data from which knowledge and conclusions have been extracted. From the initial observations recorded in a lab-book to the peer-reviewed conclusions of a journal article*".

There is a further exposition in an extended three-part piece in the 3QuarksDaily blog[2], whilst the Wikipedia entry for "open research" states:

"…the central theme of open research is to make clear accounts of the methodology, along with data and results extracted therefrom, freely available via the internet. This permits a massively distributed collaboration. Most open research is conducted in existing research groups. Primary research data are posted which can be added to/interpreted by anybody who has the necessary expertise and who can therefore join the collaborative effort. Thus the 'end product' of the project arises from many contributions rather than the effort of one group."

In 2008, Science Commons published a series of *Principles for Open Science,*[3] which covered four key elements: research literature (open access), research tools (materials such as cell lines and reagents), research data (data sets, databases and protocols) and cyberinfrastructure (open, public and extensible).

Web-Scale: This term has been used in the context of Amazon's on-demand computing infrastructure S3 and EC2[4] and by Lorcan Dempsey in his blog[5]:

"Web-scale refers to how major web presences architect systems and services to scale as use grows."

In the context of this Report, Web-scale refers to the potential for scaling up the practice of open science both within the professional community and beyond.

Open Notebook Science (ONS): This term was first used by Jean-Claude Bradley (Drexel University) in a blog post on 29 August 2006[6] where he described his novel approach to recording the details of his scientific experiments in a digital or electronic laboratory notebook (ELN). The raw data from the experiment is readable by both humans and machines in a fully transparent manner and is an example of radical sharing.

Citizen science: The Wikipedia entry states:

"Citizen science is a term used for projects or ongoing program of scientific work in which individual volunteers or networks of volunteers, many of whom may have no specific scientific training, perform or manage research-related tasks such as observation, measurement or computation….. Distributed computing ventures such as SETI@home may also be considered citizen science, even though the primary task of computation is performed by volunteers' computers".

Predictive Science: The term was used by the US National Nuclear Security Administration (NNSA) Predictive Science Academic Alliance Program (PSAAP),[7] whose stated goal was:

"to establish validated, large-scale, multidisciplinary, simulation-based "Predictive Science" as a major academic and applied research program".

In the context of this study, predictive science has been used to describe data-driven science which enables the forecasting, anticipation or prediction of specific outcomes, such as those associated with disease, behaviours or the environment.

Synthetic science: Wikipedia states that:

"Synthetic biology is a new area of biological research that combines science and engineering in order to design and build ("synthesize") novel biological functions and systems".

Earlier in 2009, scientists from the J Craig Venter Institute created a new "engineered" strain of bacteria,[8] however this practice has attracted controversy because of the claimed potential to scale-up the process from bacterial genome to the capability to create a synthetic human genome.

Immersive science: This term was used by Justin Rattner (Intel Senior Fellow), for a blog post[9] which describes development progress of a new virtual world called "ScienceSim":

"…as computing technology advances and broadband connectivity becomes ubiquitous, today's nascent virtual worlds and online games will evolve into a "3-D Internet." I believe that eventually these immersive connected experiences (as we call them) will become a primary mode for human interaction, ranging from simulated worlds used for collaboration, socialization, and entertainment to augmented realities like Google Earth that combine real-world imagery with the user-generated information".

Data informatics: The term has been used in this Report to describe library and information science methodologies and practices which have been applied to research data.

Finally in this section, it is acknowledged that the study does not seek to be comprehensive in its examination of the selected areas, and there are gaps and a degree of variation in the depth to which issues are explored. Rather, the aim has been to focus on particular perspectives with some continuing themes:

- **Team science** and the socialisation of science more generally. "*Research is increasingly done in teams across nearly all fields. Teams typically produce more frequently cited research than individuals do, and this advantage has been increasing over time. Teams now also produce the exceptionally high-impact research, even where that distinction was once the domain of solo authors*.[10]"
- **Innovation edges** and boundaries where there is maximum future potential. The term "innovation edge" was used in the title of a book by John Kao (*Innovation Nation: How America is losing its Innovation Edge….*) and as the title of the flagship NESTA Conference in 2008. It is believed that the aspects of data-intensive and open science described in this study, fulfil this description.

4 Context

Following the publication of the *Data Deluge* paper (Hey and Trefethen, 2003)[11] which accompanied the UK eScience Programme, and the *2020Science Report*[12] from Microsoft Research (2006), there is now growing acceptance of the radically changing landscape of science which is reflecting the parallel changes in distributed computing, data processing and computational analysis demonstrated by organisations such as Google. We are now in "*The Petabyte Age*" and the following statistic quoted in Wired Magazine July 2007[13] : "1Petabyte = amount of data processed by Google servers every 72 minutes", reveals the reality of this statement. In a special issue of Nature[14] focussing on "big data" and their implications for science, it is noted that "*big is a moving target*." Whilst the Web of data is not yet realised, there is a growing body of evidence which is beginning to demonstrate the potential of this vision. Research is increasingly :

- **multi-scale** – where developing data infrastructure will enable seamless and concurrent processing and integration across multiple dimensions of space, time, system and state.
- **multi-disciplinary** – innovative research will increasingly be positioned at the inter-disciplinary interface and will traverse disciplines.
- **multi-skilled** – a range of diverse competencies will be required drawn from different skill bases including computational sciences, statistics, informatics, curation / archival sciences and the domain sciences, leading to distributed multi-functional teams working together ("virtual team science").
- **multi-sectoral** – scientific collaboration and partnerships will be facilitated with organisations and individuals outside of higher education, e.g. in industry, professional bodies, learned societies, the general public.
- **multi-funded** – the multidisciplinary nature and increased scale of data-driven research, will lead to global consortia of funding bodies financing key initiatives in the emerging new digital economy.

4.1 The Data Deluge

Dramatic step changes in scale will be compounded by significant increases in complexity: the data which forms the primary foundations of scientific research will be highly heterogeneous, globally distributed and increasingly granular. Whilst this greater scale and complexity provides new and exciting computational opportunities, there are significant implications for institutions and individuals (data managers, data scientists, information scientists, librarians, researchers, funders and policy makers), alike.

In considering the relationship with science practice, a provocative essay in the July 2008 issue of Wired Magazine makes the claim "*the end of theory: the data deluge will make the scientific method obsolete*" and emphasises emerging approaches to thinking in science, based on analysis of petabytes of data rather than the traditional approach to science: "hypothesise, model, test". The article suggests that the way in which Google uses mathematical algorithms to make links, associations and correlations between massive amounts of data will lead to the traditional approach to science becoming obsolete. But not everyone agrees: "*Correlation science is pseudoscience*" (Andras Aszodi in comment on Nature online). Amongst some researchers, there is a marked resistance to data sharing where it enables secondary analysis based on data computation, which is grounded in the highly competitive environment of current research practice. There is a real requirement to incentivise data sharing to overcome barriers of this type within the research community.

Computational infrastructure is also developing to support these new collaborative approaches. As an example, the Cluster Exploratory (CluE) Programme[15] from the US National Science Foundation, is funding projects which develop research to run on a large-scale distributed computing platform developed by Google and IBM in conjunction with six pilot universities: "*bringing cloud computing to academia*". The cluster will consist of approximately 1,600 processors with open source software and includes participation by the University of Washington, where the cluster is used as part of the undergraduate curriculum in Computer Science. In a similar manner, Cloudera[16] is now providing access to Apache Hadoop, an open

source implementation of MapReduce, Google's software framework for the deep analysis and transformation of very large datasets. The Dryad infrastructure from Microsoft Research, is also providing access to computer clusters or data centres, as an approach to addressing issues of parallel programming at scale.

The prospect of widespread and routine processing of very large datasets has significant implications for the provision of a robust and resilient data infrastructure with supporting data curation and preservation functionality. A number of reports have highlighted a set of Stewardship Principles (RIN)[17], roles, rights, responsibilities and relationships for data curation (*Dealing with Data Report*)[18] and cost estimates for a proposed UK data service (UKRDS)[19]. Whilst there is a growing body of good practice accumulating, the rapid pace of change makes co-ordinated strategic planning and informed policy development for research data management more crucial than ever before.

4.2 The Socialisation of Science

The viral expansion of highly social and participative computing models, tools and services provides really exciting possibilities for complementary approaches in the future. The embedding of social networking sites in people's day-to-day lives has crossed over into the professional realm, leading to these tools being applied in new contexts. Some of these social networking platforms are noted in Section 6.1, and we can begin to see how the socialisation of science is gaining traction and changing practice at the research coalface.

The explosion in user-generated content (images, opinion, ratings, memories) framed variously as posts, feeds, streams, 'casts and tweets, suggests that there is a very active, willing and able population who can provide potential capacity and capability to contribute and curate the burgeoning volumes of data on the Web. Both the massively parallel computation-centric and participative crowd-centric models allow us to tackle societal and computational problems and research challenges, which are beyond the capacity of a single individual or machine, or one research group or discrete network.

The shift towards "team science" has been noted: by tracking citation rates, Brian Uzzi was able to conclude that "*there's something about between-school collaboration that's associated with the production of better science*" Team science has been discussed in the setting of particular disciplines, such as psychology,[20] in the context of interdisciplinary research[21] and summarised in a Nature article[22]. The development of Virtual Research Environments[23] and other collaborative infrastructure, has provided platforms for distributed research groups to work more effectively together.

This Report extends the concept of team science by exploring aspects of open (participatory) science, including wider professional collaborations and public participation.

4.3 Ethical Concerns

The trend towards increasing openness brings with it significant challenges in dealing with moral, ethical and philosophical issues. For example, 23andMe[24] based in California, offers a kit which can be purchased to enable submission of a saliva sample to the Lab for analysis and subsequent "*exploration of your genome*". Data encryption allows you to share as much or as little of the data as you wish, with your family, friends and beyond, facilitating identification of inherited genes and associated health information. Whilst some US states have prohibited direct-to-consumer genetic testing, the company has obtained licences to continue trading in California and 23andMe's DNA testing service was awarded *Invention of the Year* by Time Magazine in 2008, for pioneering retail genomics.

Synthetic biology, where biological engineers build biological parts which may be assembled to construct new components and new systems, is another case where significant intellectual property rights, ethical issues and philosophical questions arise. These strings of DNA bases can be compared to software source code and a number of legal, patent and licensing issues soon surface. The notion of a synthetic biology commons[25] has been proposed linked to the

BioBricks Foundation, which has been created by scientists at MIT who are working with the Registry of Standard Biological Parts.

4.4 Barriers

This Report refers to innovation edges, but it is important to note that there are also barriers which are impeding progress, restricting the creation of new knowledge and potentially resulting in poor return on investment of public funds. Some of the barriers are tangible and concrete but others are cultural and more challenging for their opacity. "*Data sharing: Empty archives*" was the stark headline of a recent Nature article[26] which explored why many researchers are not depositing their data in (institutional) repositories for open sharing. Whilst there are areas where communities have been successful in developing a sharing culture (the arXiv.org community is one example), "*these discipline-specific successes are the exception rather than the rule in science*."

4.5 Perceived Value and Benefits

At this early stage in the Report, it will be useful to articulate the perceived value and benefits of open science**,** whilst acknowledging that further evidence is required to support these assertions and that any cumulated benefits will take some years to become fully apparent:

1. **Increased return on investment of public funds** allocated to science and research through making data outputs openly available for re-use.
2. **Faster dissemination of research outputs** including methodologies, data, models and scientific outcomes.
3. **Greater academic rigour**, robustness and scholarly integrity from transparent data practices.
4. **Higher potential for new discoveries** and new knowledge arising from data re-use contributing to growth in UK economic and intellectual wealth.
5. **Accelerated ability to predict scientific outcomes** and behaviours based on large-scale open data analysis, shared complex models and simulations.
6. **Efficiency gains** from open research practice leading to reduced unnecessary repetition of research activity and associated wasteful funding allocations.
7. **Enhanced opportunities for student learning** from open sharing of experimental methods and results data.
8. **Increased human capacity and capability** from professionals, amateurs, volunteers and citizens to assist in collecting, curating and preserving the growing scientific record.
9. **Enhanced public engagement and understanding of science** principles and practice through raised awareness, pro-active participation and direct contribution to research.
10. **Significant wider societal gains** through more inclusive and participatory approaches which facilitate public empowerment and ownership of global challenges.

5 Scale, Complexity and Predictive Potential

In recent years, the continued growth in computational power, processing speeds and data modelling, has led to an increasing ability to tackle significant societal and scientific challenges. Radically new approaches, asking new questions with new conceptual and technological tools is leading to nothing less than a scientific revolution. We are able to apply state-of-the-art computational tools to living systems. For example, we are able to address fundamental biological research problems at molecular and cellular levels, at the organism level, and now increasingly at a holistic systems level where inter-disciplinary data, sophisticated analytical and

modelling techniques may be integrated and combined to simulate and model complex biological processes.

In addition, modern research and computational hardware, software and techniques such as high-throughput processors, enable scientists to perform experiments and calculations at different orders of magnitude of scale and volume: research that was completed in a year can now be repeated in a weekend. “Next-next generation” sequencing technologies will facilitate the emergent area of meta-genomics[27], which gives evidence of the dramatic scaling up of science practice. Pacific Biosciences claim that single molecule sequencing technology will provide a complete human genome in 15 minutes by the year 2013[28].

There are many practical issues around working effectively at multi-scale. A new JISC-funded project Infrastructure for Integration in the Structural Sciences (I2S2)[29], is examining the challenges faced by scientists working in the chemistry domain who are using a mix of centralised large-scale facilities (such as the DIAMOND Synchotron at the Rutherford Appleton Laboratory in the UK), and local small-scale laboratory facilities. Researchers need to move data across both institutional and domain boundaries, in a seamless and integrated manner: I2S2 seeks to “bridge the chasm” and develop a robust data infrastructure to enable these seamless transformations.

The harmonisation of distributed representations of data models through abstraction into an Integrated Information Model, which underpins the research across multiple sites and global locations, will be a significant step forward within the structural science community. The Information Model will be built on an amalgamation of solid foundations already established, namely the Core Scientific MetaData (CSMD) Schema[30] developed by the Science & Technology Facilities Council (STFC), and the Curation Lifecycle Model[31] developed by the Digital Curation Centre (DCC). It is hoped that the extended CSMD will become the core information model to support structural science across the experimental lifecycle, interoperating between large facilities and laboratory based science.

5.1 Data Modelling

We are also seeing sustained growth in data modelling, both in terms of the scale of the models, the complexity of the simulations and the associated visualisation requirements to enable interpretation and analysis by both humans and machines. Examples include the mathematical modelling of pandemics such as influenza and SARS (Imperial College), earth system and biosphere modelling such as forest dynamics and other related ecosystems (Microsoft Research), climate modelling (Climateprediction.net) and life systems modelling such as transscaler 4D modelling of pancreatic organogenesis in the mouse (Microsoft Research).

In this latter instance, spatial changes in three dimensions and a fourth dimension time, are visually mapped. The modelling facilitates exploration of *in vivo* developments *in silico.* In the virtual pancreas, the computational model anticipates the *in vivo* biological experimental findings: the *in silico* visualisations of morphogenesis reveal a close resemblance to histological images of the pancreas[32]. In this study, the authors note four aspects of integration achieved by the simulation and modelling experiments:

- *Multiscale – Four dimensionality enables concurrent interactions (genetic, molecular, intracellular, environmental etc.) to be visually comprehensible.*
- *Cross-scale – Facilitates understanding of the interplay between organ structure and environment with molecular interactions, and vice versa.*
- *Emergence – Four dimensionality discloses the emergence of new properties across scales.*
- *Dynamics – Changes with time become visually perceptible.*

This type of modelling which embeds static experimental data into a reactive model linked to a 3D animated front-end for visualisation and a mathematical interface for analysis, also allows new experimental questions to be asked and new experiments to be carried out, thus saving the researcher time and giving focus and direction to the research. If this level of life system

modelling is extended, we can envisage the ability to link and integrate computational models, to model pathways across multiple systems and to design *in silico* organs or organisms: we will be moving towards the virtual human[33].

5.2 Data Visualisation

One key element common to all modelling and simulation work is the importance of data visualisation techniques and associated skills, illustrated by the 3D molecules and NMR spectra from ONS experiments, where the molecular stereochemistry can be visualised in virtual worlds such as Second Life on Drexel Island. This and many other examples of compelling data visualisations are presented in the recent monograph "*Beautiful Data*"[34]. A useful overview of Web 2.0 data visualisation tools is presented in a two-part study from the JISC UK Datashare Project[35]. Given that the average researcher is usually time poor, "glanceability": (enabling users to understand information with low cognitive effort[36]) acquires increasing importance.

5.3 Predictions and Forecasts

The ability to simulate and model aspects of our world and life itself, leads us to begin to be able to make predictions based on these models. In 2008, the US National Nuclear Security Administration (NNSA) announced a Predictive Science Academic Alliance Programme (PSAAP) with five universities leading large-scale multi-disciplinary simulation-based initiatives which cover areas such as materials science (predicting impact behaviour of projectiles), aerospace radiation challenges, hypersonic flight and space science. These types of advanced simulation enable scientists to ask challenging "What if?" questions, to model extreme scenarios and to quantitatively assess the outcomes, the likelihood, the significance and the risks. It is clear that such developments could have a beneficial impact on other areas of research such as immuno-therapeutics and predictive medicine e.g. the use of biometric data to predict disease as noted in the UK Digital Curation Centre SCARP Project Neuro-imaging in Psychiatry Case Study Report[37].

In a 21st Century research environment, where the absolute reliance on exascale resilience of HPC systems[38] and on robust distributed data infrastructure is paramount, guaranteed standards of data quality ("reference datasets") together with authoritative validation and verification of derived simulations and models, are essential. Aside from the observed requirement for more sophisticated mathematical input to develop complex models at extreme Web scale, there are significant challenges linked to understanding the relationships between theoretical and static models of dynamic systems, and with real-time and continuous modelling, where the dynamic nature of the system models results in constantly changing data outputs.

This challenge is illustrated by the European Coastal Sea Operational Observing and Forecasting (eCOOP) System where 72 partners are working on over forty models and data feeds from at least 15 institutes to compare and overlay observational data with the model simulation results. Software is used to track and forecast storms and confidence in the underlying models is key: eCOOP data visualisations are available from the Reading eScience Centre Godiva2 Project demo pages[39]. There are issues associated with determining and certifying data quality, which in turn affects the calibration, validation and verification (confidence) of the derived model. The assessment of the accuracy and robustness of a prediction is vital; this is linked to uncertainty quantification, which may be dependant on the accuracy of the underlying model, and the integrity of the data from which the model is derived. There are also issues around the wider sharing of models (rather than the underlying data), associated standards (e.g. Minimal Information Requested in the Annotation of Models MIRIAM[40]), workflow/visualisation sharing tools (such as myExperiment, Taverna and Utopia) and model repositories (such as Biomodels.net), to quote examples from the bio-informatics domain.

In another example, the GoogleFlu estimates and predictions of weekly influenza outbreaks in regions of the United States, show how search query data and collective intelligence act as indirect signals of disease outbreaks and health trends, and which pre-empted "official" surveillance reports based on clinical and virus data from the Centre for Disease Control, by 1-2

weeks[41]. The derivation of models from the processing of billions of individual searches from five years of Google search logs, demonstrates an innovative and socially valuable application of Web data.

Clearly, the accessibility of openly available datasets, (both established reference datasets such as the Protein Data Bank and less mature data collections), will be crucial in facilitating the development of robust predictive models and simulations, which enable the forecasting of environmental events, health and disease, behaviours and future markets. It is also important to recognise that the increasing dependency on complex and dynamic models and frameworks, leads to specific challenges associated with the curation and long term preservation of these derived models and simulations, in addition to the challenges of curating and preserving the underlying data.

Consultation Challenge 1: Scale, Complexity and Predictive Potential

Some key questions address community awareness and understanding of the potential implications and impact of (open) data-intensive science at new extremes of scale and complexity, and the service requirements for associated data curation and preservation.

What is the level of awareness and understanding in the wider community of the prospects and societal implications of predictive science?

How are the methodologies and tools for data quality, validation and verification, which underpin robust and trustworthy large-scale models and simulations, implemented in different disciplines? Are appropriate data quality standards in place?

How are the necessary mathematical skills available to science teams, particularly in domains such as biology?

How can services like the Digital Curation Centre, best support the effective curation and long-term preservation of complex and dynamic data models, simulations and visualisations?

6 Continuum of Openness

One of the definitions of open science cited in Section 3 referenced the life sciences but of course, open science principles may apply to the full trans-disciplinary spectrum and not just to pure and applied science. Michael Neilsen refers to “extreme openness” [42] and we can consider a “**continuum of openness**” which transitions from so-called “dark data” perhaps protected by the lone scholar through incremental stages towards a fully openly accessible, shared and participative environment with public contributions (both voluntary and paid) at the other extreme; this is illustrated with selected examples in Figure 1. The ability to make “dark data” available i.e. data from failed experiments[43] is causing disquiet amongst some scientists and clearly it would be inappropriate to disclose certain datasets in this way. The concept of a continuum of openness is one way of acknowledging that a range of publishing channels is required. The boundaries and social relationships between familiar concepts of openness, sharing, curation and collaboration are subtle. The economics and politics of science require recognition of the tensions between collaboration and competition (Collaborate to compete).

In parallel, increasing scale leads to requirements for greater automation and machine processing of tasks. Open data enables loosely-coupled collaboration in addition to more formal consortial partnerships. To take this a step further, if the Continuum in Figure 1 showed a third dimension, the “z” axis might be a cognitive processing transition from human to human (h2h) data interaction and exchange, through to humans using information and computer technologies to facilitate data capture, processing and sharing (h2m), to wholly machine to machine (m2m) transactions, the latter illustrated by sensors capturing and streaming data which is then routinely processed, analysed and submitted to further automated workflows.

In general, the perceived barriers to openness are not technical but are overwhelmingly cultural, social and political in nature. Some of the barriers to data-sharing are described in a presentation by Heather Piowar[44] which also includes selected survey results, for example "*80% scientists report positive experiences from data-sharing*". Science today is highly competitive and a mindset change is required to promulgate more collaborative and participative approaches. This can be assisted by the funding regime which is fundamentally a competitive process, however some funding agencies are promoting open policies such as the Wellcome Trust and the National Institutes of Health in the US. Intellectual property (IPR) issues related to commercial exploitation may also act as a barrier to data-sharing, but sometimes this particular barrier is of more importance to the institution than to the scientist.

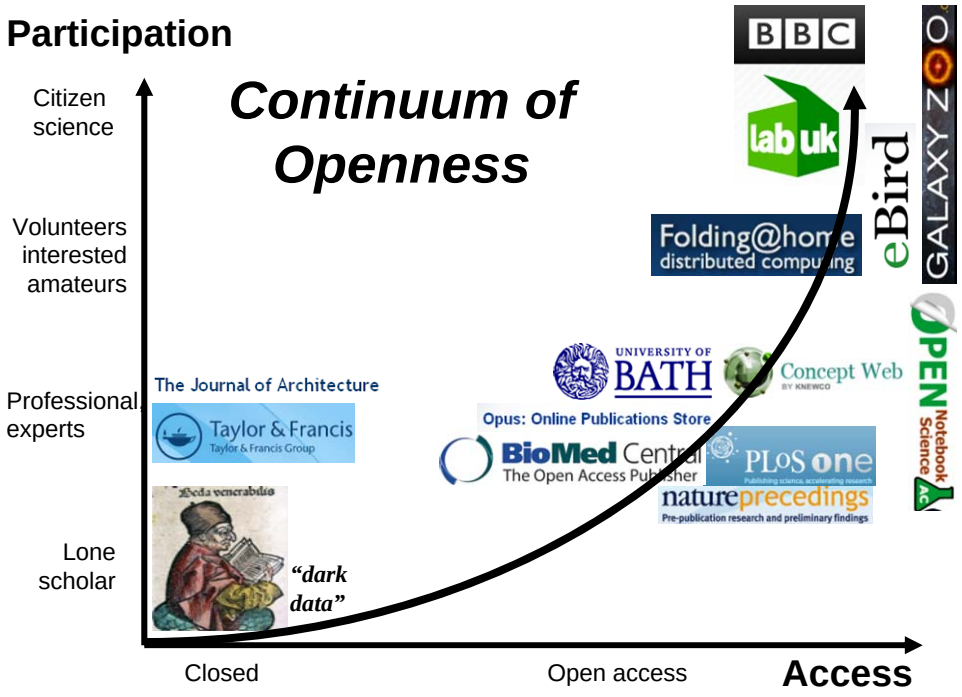

Figure 1 Continuum of Openness

6.1 Social Tools and Platforms

The availability of good social tools and platforms is crucial to the growth of open science. Whilst there are a range of tools in use, there is scope for an in-depth analysis of the functionality and benefits of existing tools with requirements for further development. Currently, researchers are using open science tools such as:

- **Connotea** for reference management
- **Mendeley** (which applies LastFM principles associated with music selections to journal articles)
- **Friendfeed** (for threaded discussion and aggregation)
- **Scivee** and **YouTube** (for sharing experimental methodologies and protocols)
- **SciLink** and **Nature Networks** (for social networking)
- **myExperiment** (for sharing workflows)
- **eyeLIMS** (an open source Laboratory Information Management System)
- **LabLit.com** (about science/laboratory culture in the literature and media)

- **ConceptWeb** (from WikiProfessional and includes WikiPeople and WikiProteins) .

A preliminary list of open science tools[45] has been produced with categories such as "*blog collections, blog aggregators, social networks, protocol sharing and literature sharing*" supplemented with a list of critical criteria such as "stability, architecture, contextualisation, design and features". Further examples of open science practice are described in the next sections. Whilst some tools are relatively well-established, there is interest in the potential scholarly applications of Google Wave[46].

One other new service to support open science is InkSpotScience, which was developed by academics from the University of Newcastle, and provides an active workbench for scientists as an alternative to proprietary and open source tools, allowing them to work at home e.g. on drug discovery, more effectively. The service provides infrastructure including secure hosting for Web services, on demand computing with support for scripting and workflows and digital signing. The service is free for open science applications.

6.2 Blogs and Blogging

The current scholarly publishing model focussed on peer-reviewed articles in subscription journals, has been in place for centuries, however there are now other (more open) publishing channels on the Web, such as blogs. The scepticism and slowness of scientists to embrace blogging has been noted: "*Scientists don't blog because they get no credit for that*" (Chris Surridge, PLoS ONE) and "*its so antithetical to the way scientists are trained*" (Huntington F Willard, Duke University).[47] Nevertheless, a number of events have explored scientists' experiences and attitudes to blogging, including Science Online 2009[48]. Sites such as Research Blogging provide a forum for the online discussion of peer-reviewed research, Chemical blogspace blog synthesises posts from many blogs in a digest format, whilst *The Open Laboratory* is an annual anthology of the best science writing in blogs, (rather ironically) published in book format[49].

A range of issues have been associated with why scientists have been relatively reluctant to adopt these social software tools and some of these are briefly listed below:

- **Vulnerability to data predators and "scooping":** a particular problem in certain fields where the time lag between discovery and validation / publication is relatively long. Questions arise around how much information do you post? Do you share the underlying model? Can a blog or wiki posting be accepted as proof of priority for a patent? These types of concern associated with credit and attribution may have far-reaching impact on the careers and tenure of academics. If scientists have tenure then there is less associated risk, but new-entrant scientists may have different views.
- **Trust, quality and peer review issues**: the more journalistic and opinion-piece type of content in blogs do not carry the imprimatur of formal peer review mechanisms.
- **Time constraints**: the perceived time spent away from the lab bench. Time may be spent looking rather than doing: blogging is seen as a secondary activity.

On the more positive side, there are advantages too:

- **Integration in the scholarly business model**: for some publishers such as the Nature Publishing Group, blogging is part of the business model. Nature runs many blogs and has sponsored blogging workshops. Blogs enable communication between scientists and society at large and contribute to the core mission. Blogging is an integrated part of the science process and is complementary to the published literature, which may be viewed as the gold standard. Adding value is seen as key: blogging about the published literature is embedded as part of the scientific record.
- **Teaching***:* a blog can be used to share methodologies, to obtain feedback, to engage with students and to acquire help and advice on technical topics.
- **Outreach channels to other disciplines and the public**: when you want to simply "share stuff". Blogs can be used to comment and discuss methodologies, results and research outcomes.

- **Find collaborators**: Blogs can be an effective route to make connections with scientists working in a similar area (the social phase of social software). If key documents are openly shared, alerts can be triggered when documents are updated, directly providing links to potential collaborators.
- **Grant proposals**: Blogs can be used to share ideas and gather support for prospective grant proposals and can also be used to disseminate information about successful grants, new projects and initiatives in order to maximise impact.
- **Recruitment and Employment**: Blogs can be used as recruitment channels. Bora Zivkovic posted in his *Blog Around the Clock* for a job at PLoS and was supported by posts from colleagues. A blog can contribute to a student's e-Portfolio and help to provide skills evidence for the student when they seek employment. Social Web evidence may be useful for the employer to estimate the qualities of the prospective student employee.
- **Reputation:** Blogging can be a highly effective channel for raising personal visibility and profile: it is possible to achieve a degree of fame based an overtly open approach.

6.3 Peer Production

The proliferation of social Web tools has facilitated a more collaborative approach to both big and small science, enabling globally distributed teams to work together, share data and documents, discuss experiments and publish results. Scientists collaborate to pro-actively curate large community / reference data-sets, performing data cleansing, annotation and other management tasks. The scientists, (within project consortium partners or members of the wider disciplinary / inter-disciplinary community), work together to achieve common goals. There is added value in harnessing professional and expert community resources in this way, since trust in the quality of the data, is a major factor in assuring its citation and subsequent re-use.

In the genomics and post-genomics domain, professional community curation with associated social infrastructure is essential to ramp up the annotation effort to match the scale of data generation[50]. WikiProteins[51] and WikiPathways[52], provide examples of community curation where established experts and students may work together to edit, update and maintain the data pages supported by "bots" which identify areas where there are inconsistencies, redundancies and incomplete data. In a new collaborative effort in bioinformatics, the Concept Web Alliance[53] is a non-profit organisation bringing together social Web and semantic Web initially in the life sciences, and has been supported by the Netherlands Bioinformatics Centre.

Concept Web Alliance Mission:

"To enable an open collaborative environment to jointly address the challenges associated with high volume scholarly and professional data production, storage, interoperability and analyses for knowledge discovery."

A further example of collaborative action is the ChemSpider service now hosted by the Royal Society of Chemistry, which provides open access to millions of chemical structures and supports community curation as a means of "cleaning up" the data and so increasing the quality and accuracy of the content. A Curation Manual has been published[54], which includes guidance on the maintenance of identifiers, links to Wikipedia articles and the deposit of new structures. ChemMantis is the ChemSpider Journal of Chemistry: an experimental open science journal.

6.4 Open Notebook Science

Cory Doctorow has quoted the transition from alchemy to chemistry as an exemplar where changes in behaviour i.e. a new culture of sharing, transformed the domain[55]. In chemistry today, there are a growing body of projects, services and initiatives promoting the open dissemination and re-use of datasets; a paper by Peter Murray-Rust[56] outlines the basic concepts. This prior work includes eBank and eCrystals Projects (federating repositories for institutional crystallographic datasets), SPECTRa (deposition and validation of primary chemistry research data), SPECTRa-T (data in theses), R4L (repositories and blogs in the

laboratory), the Australian TARDIS initiative (sharing raw X-ray diffraction data) and CombeChem/SmartTea (smart laboratories) initiatives.

One area where highly innovative open development is taking place is in the laboratory, where methodologies, protocols, materials, environmental conditions, experimental results and conclusions are recorded in (relatively) informal note form in laboratory notebooks, which are traditionally paper-based and positioned as a "diary" of experiments carried out by the bench scientist. Two Open Notebook Science (ONS) examples are outlined here.

6.4.1 UsefulChem

The UsefulChem wikispace[57] has been developed by Jean-Claude Bradley (Drexel University) and his team, for research and undergraduate teaching and is chronicled in the Useful Chemistry blog. In the research context, a pioneering paper in Nature Precedings[58] describes the underlying platform used in the determination of the Ugi reaction. ONS provides additional valuable data and procedural information where conclusions are fully supported by evidence, which may supplement established peer review mechanisms. The examination of failed experiments is particularly useful: because the raw data is available, data outliers can be identified and tagged "do not use" together with a reason.

Detailed interactions with students are possible with visual evidence and comments recorded in text and histories, providing excellent learning opportunities during the scientific apprenticeship period. An audit of an experimental process can be carried out to check procedural detail: edits and deletions are recorded so a full log is available to describe the laboratory procedure. An alert system via RSS or email notifies users of changes. The recording of temporal changes is crucial and a third party timestamp service is used on wikispaces to help to manage versioning issues. Service resilience is enhanced through implementation of an effective back-up policy, however there are currency and replication challenges given the real-time nature of the platform and a key issue is the balance between replication versus redundancy.

A laboratory notebook is most useful when linked to other key sources and tools. These include a blog for discussion, Googledocs for collating results and comparing experiments, Flickr for hosting images of experiments, a "scribble" tool for annotations with subjective comments from scientists highlighting which data is of interest, supplier catalogues for empirical data about chemicals used and RFID tags for recording physical samples and linking to digital records. Development of an underlying data model and schema has begun with the objective of defining standard terms such as ADD, VOL to facilitate machine-readable formats. Links with myExperiment are being pursued in this context. InChI and SMILES codes identify specific molecules which can be searched with Google.

6.4.2 ChemTools

A different approach has been adopted by Jeremy Frey's team at the University of Southampton with a series of blog-based ChemTools[59]. A blog platform is used as a laboratory notebook and a series of blogs capture output from laboratory instruments, machines and sensors (room, door, light, temperature etc.) as well as from the scientists themselves. ChemTools has two stated aims: firstly to act as a day-to-day laboratory record and secondly to facilitate machine processing of the experimental data to enable more sophisticated information to be extracted. Data feeds from instruments can be aggregated, published on the open Web (so that the outputs can be viewed remotely) and mashed / re-used. Each data item (posts, samples, procedures, products) has a post with an identifier, creating a linked network of posts. The CLARION project at the University of Cambridge, is implementing a commercial Electronic Lab Notebook (ELN) system and will publish open data with additional semantic definition through use of RDF (Resource Description Framework) and CML (Chemical Markup Language).

6.4.3 Future Development and Implementation

A variety of issues have arisen from the ongoing development of ELNs concerned with functionality of the tools, their usability and the sustainability of the content and they are summarised briefly here. One basic practical issue noted with text-driven Web services such as

blogs, is the management and manipulation of tables, where a template and consistent tags and metadata are needed to streamline editing, support standard notation and facilitate data extraction. There is continuing discussion about developing and applying data models to ELNs; the Functional Genomics Experiment (FuGE) model is just one option that has been postulated.

The application of persistent identifiers such as OpenIDs and DOIs, in open science more widely relates to issues of identity management, timestamps, versioning, tracking, provenance and citation, which continue to provide challenges to ONS approaches. A different issue is around visualisation of the many posts and the Timeline tool from the MIT SIMILE Project provides useful functionality in this context. There are issues around legacy hardware requiring scanning of paper-based information into the ONS system, issues around the provision of safety information within the ONS lab, questions around the degree of openness of the notebooks, discussions about requirements for electronic signatures (of the scientist and their supervisor), use of data licences and institutional policy regarding ELNs. The ease-of-use of ELNs has been noted with the caveat that user education and some degree of computer literacy is required, even with these relatively non-specialist Web tools. Some significant challenges are associated with the sustainability of ONS approaches, the archiving and curation of ELN content and long-term preservation of ELNs as part of the scientific record.

One view of the laboratory record of the future is given by Cameron Neylon[60], who describes the "linked data Web native Lab Notebook" as a "semantic Web ready" laboratory record. In this Linked Data world, self-describing data files[61] would be connected to related datasets, blogs, wikis, Web services such as ChemSpider, repository papers and the peer-reviewed literature. The longer term prospect for ONS is still open to question. Supporters contend that the process is ultimately significantly faster in the dissemination of results and the outcomes of funded research. However its real value will become apparent when a critical mass of research is conducted, monitored, described and shared in this open fashion.

Consultation Challenge 2: Continuum of Openness

The following questions address the awareness and understanding of fundamental open science concepts and are supplemented by probing exploration of practitioner experience.

What are the views of the community on open science principles, acknowledging that "openness" is a continuum or sliding scale with different groups, services, information and data, positioned at different points?

What are the views of the community on the perceived value and benefits of open science methodologies? How can these benefits be demonstrated and evaluated?

Should research funding bodies be pro-actively supporting open science principles and practice? What are the policy implications? What infrastructure is required?

How aware are the majority of scientists of the range of social Web tools available to support open science? How are the tools used in different disciplines? What are the perceived advantages and disadvantages of using collaborative tools? How can social tools add value to research? What are the cost-benefits of using these types of tools?

What are the implications of open science communication channels e.g. blogs, on scholarly publishing models? What are the views of publishers and learned societies?

How can the peer production model for data curation, be applied and adopted in other disciplines?

What are the community views on Open Notebook Science? Should these radical methods be migrated across to other disciplines and if so, which other disciplines would benefit? What key ONS development and enhancement issues need to be addressed?

7 Citizen Science

In Section 6, the concept of a Continuum of Openness was proposed where the level of data sharing varies from the closed dark data which is never made available, to the fully open and public datasets on community Web sites like Swivel. There is however another perspective on openness to consider: the scope and type of participation and contribution.

7.1 Engaging the Public in Science

If we take the participation theme a step further and extend the science team to include interested volunteers or amateur scientists or citizens, then some very exciting opportunities emerge. In some domains, citizen science has a long history; consider the Victorian naturalists and areas of ornithology (e.g. National Audubon Society Christmas Bird Count which has taken place annually for over 100 years), astronomy, meteorology and archaeology, where an emphasis on observational recording was central to the science and to the scholarship. We are now seeing a veritable resurgence in citizen science with the social culture of the Web beginning to influence and radically change the way science is performed. The announcement of the formation of the Citizen Cyberscience Centre, a collaboration between CERN, UNITAR and UNIGE, is a strong indication of the perceived importance of this approach, particularly for international collaboration, for developing countries and for neglected diseases.

A mature open science example is GalaxyZoo, which has developed a community of amateur (armchair) astronomers who collectively help to classify galaxies via customised user interfaces, successfully combining human observational and pattern recognition capacity with categorisation capability. The public work alongside disciplinary experts in a truly global initiative to help to collaboratively map the universe. Recruitment to the international GalaxyZoo team in 2008 resulted in the advertisement of two postdoctoral research posts in “Internet-based Citizen Science” at the University of Oxford, working in the Department of Physics.

In a further example, the BBC LabUK Initiative[62] is harnessing community effort in online experiments and is seeking to work with scientists to help to solve professional research challenges which are suited to the type of mass participation which can be achieved through this medium. Exemplars such as BBC SpringWatch, eBird[63] and Bioblitz Bristol have harnessed Web and mobile technologies to engage the public in collecting natural history data and this approach is particularly effective for monitoring species living close to humans.

7.2 Learning from Citizen Journalism

Using cameras in mobile phones to provide real-time images and video, has parallels with the growth in citizen journalism. In a blog post entitled “*How citizen journalists can learn from work of “citizen scientists”*, Dan Schultz[64] outlines three classes of scientists: professionals (who make a living from science), amateurs (who tackle science as a hobby) and citizens (equipped to contribute to science when they are empowered by tools and networks). Schultz further explores the types of role and tasks that members of each group might adopt in working symbiotically, and begins to consider aspects of authority, supervision, standards and best practice.

The lowering of barriers to participation (most people have a mobile phone), together with the developing potential for automated metadata generation (geo-spatial, time etc.) offer great opportunities to gather environmental data about the world around us. In the future, we can expect the further development of sensor-rich mobile devices which include geospatial mapping using GPS, together with embedded sensors which might measure ambient temperature, to effectively deliver a mobile sensory network of environmental data and associated metadata for public census work and participative surveys: **participatory urbanism**[65].

In one exemplar, the EpiCollect Project[66] has developed a generic framework currently based on Android phones (but with support for other operating systems coming on stream), for community data collection in epidemiology and ecology. Figure 2 shows a simple field test using soil samples across Southern England by a single field worker with EpiCollect data submitted to

a project Website. Clicking on a sample point displays the variables collected, such as soil pH, temperature, moisture, GPS position, date etc. and any photograph associated with the record. Communication between the project curator and the field worker is via Google Talk instant messaging.

However, the monitoring of human behaviour using mobile devices e.g. life-logging and similar initiatives, raises privacy and legislative issues: the Google Street View service is a case in point where security issues have been contrasted with the wider benefits for populace as a whole. Scott McNealy, CEO Sun Microsystems said famously in 1999, "*You have zero privacy anyway. Get over it*", but there is a requirement to strike a balance between social good and privacy. Concerns remain, but the benefits to science, public health and society, may strengthen arguments for these types of service.

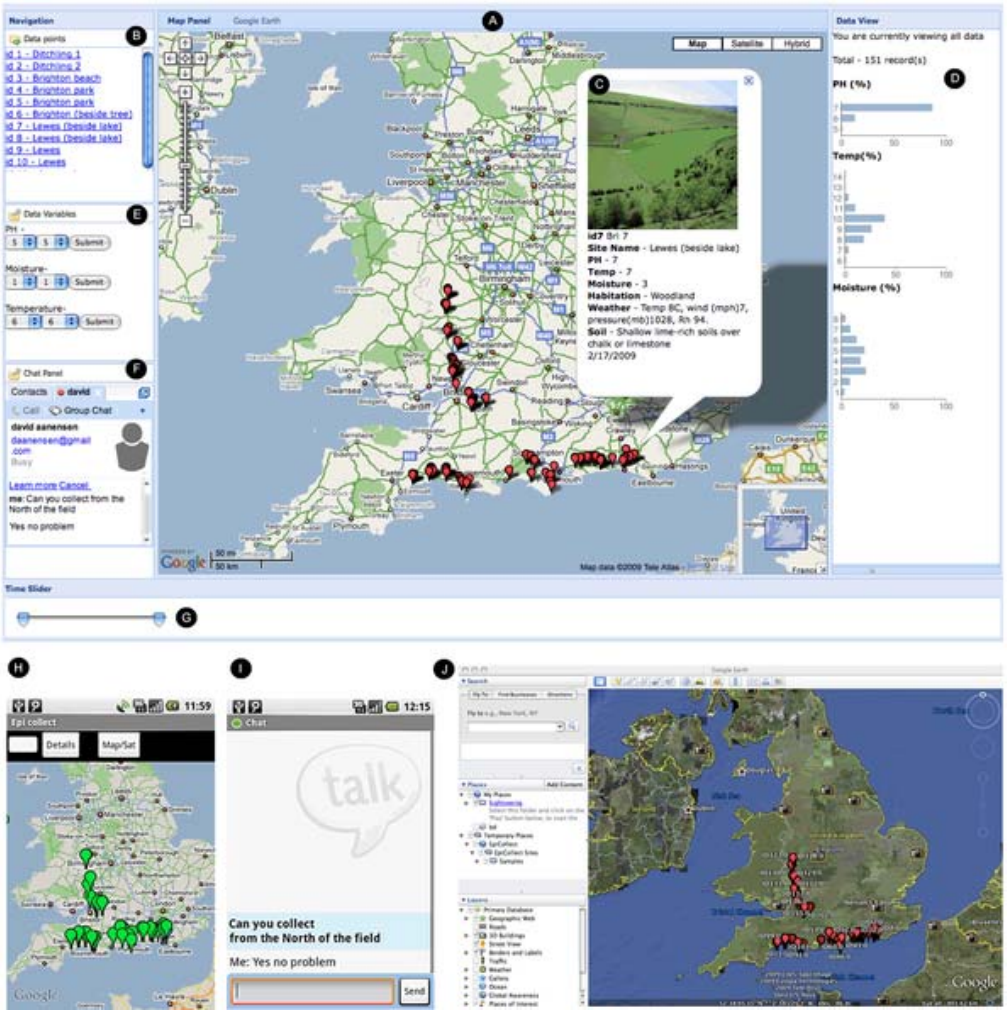

Figure 2 EpiCollect : simple field test (Reproduced from PLoS ONE)

7.3 Volunteer Computing

In a different approach to harnessing community resources, a series of volunteer computing initiatives based on the open source BOINC software from the University of California at Berkeley[67], has utilised spare computer processing cycles on public computers to provide additional computational capacity for scientific analysis. One of these initiatives Rosetta@home, which is determining the 3-dimensional shapes of proteins, has taken the level of participation a

stage further by drawing on the principles of computer gaming. The interactive Foldit[68] game allows contributors to "*solve puzzles for science*" by taking advantage of human puzzle-solving ability with people playing competitively to fold the best proteins. Foldit attempts to predict protein structures; future developments will add functionality to the game to allow users to design new proteins that could help prevent or treat important diseases.

Rosetta@home also supports individuals and teams (which may be departmental, institutional or national), and has a competitive credit or points scoring system to record contributions, with a league of top participants. These types of competitive element are found in many multiplayer games, where the more mundane aspects of the game are framed and manipulated around a series of levels giving a sense of achievement to the player, which helps to progress the game forward. Any task may be reformulated as play or as a part of a game (Jane McGonigal): Alternate Reality Games such as World Without Oil, allow distributed players to work together to collectively explore future scenarios. There is great potential for embedding this type of approach in the scientific process, but currently, these concepts are in their infancy.

7.4 Service Design and Development

Designing services for citizen science or developing services where experts and non-experts work together, is challenging. There are many interface design and usability issues associated with the different levels of knowledge, expectation and need. Tailored interfaces are required for successful public interaction and engagement: for example relatively simple forms with good Q&A are essential for effectively collecting observations. In addition to front-end presentational systems, there must be annotation authentication systems, harvesting mechanisms for gathering tags and processes for combining authoritative metadata with user-generated social tags, which are then fed back into machine learning systems. A candidate architecture for harvesting and aggregating networked annotations has been developed in the HarVANA Project. [69]

7.5 Harnessing Cognitive Surplus

Whilst BOINC initiatives capture redundant computer cycles, the reCAPTCHA[70] initiative seeks to make use of "wasted human processing power" or "human computation", to improve the digitisation process. Where the standard random CAPTCHA displays are used to differentiate between humans and computers in many Web transactions, the reCAPTCHA displays are words from scanned texts which OCR programs are unable to recognise.

The application of crowd-sourcing approaches to research challenges has had some success in initiatives such as the ONS Solubility Challenge[71] where the community was called upon to help to measure the solubility of compounds in organic solvents in an initiative sponsored by the chemical company Aldrich, Submeta and the Nature Publishing Group. The winner of the first ONS Submeta award was a chemistry undergraduate student, and more awards will be made during 2009. Whilst this exemplar used the freely available resources of other "experts", there are cases where ideas and effort have been harnessed from the wider community e.g. the Open Prosthetics Project, where a mix of users of prosthetic devices, experts and funders collaborate to innovate and improve device design and implementation.

A number of corporate organisations such as IBM, Dell, Pfizer and Proctor & Gamble have drawn on the creative thinking and ideas of the wider population of Web users to help to solve Grand Challenge problems: *"the smartest people never work for you"*. This approach has been extended into a business by sites such as Innocentive, where you can register a challenge in a range of categories including life sciences and assign a cash reward. As an example, "*Novel inhibitors of proteases and lipases. is a Reduction-to-Practice Challenge that requires a written proposal and experimental proof-of-concept data*". Amazon Mechanical Turk, described as "*a marketplace for work*", is a similar business instance, but one where money is exchanged.

7.6 Changing Business Models

In business terms, each of these exemplars may be characterised as a "**collaboration market**"[72] where ideas, questions, data and resources are exchanged. They illustrate how a collaboration market can drive innovation based on new models of trade, and we can begin to see how these approaches may influence and radically change the traditional scholarly research environment. The combination of spare brain capacity, spare computer cycles and time, which when aggregated achieve critical mass, means that transaction costs drop radically: the "cognitive surplus" proposed by Clay Shirky. We can begin to speculate on what intellectual assets may be traded or exchanged in an open science environment and which may not be shared: for example scientists may share their raw data but not the models derived from these data, and there may be disciplinary differences in these judgements.

Further indicators of radical changes in science business models are given by the pharmaceutical company Merck, which has pledged to donate a significant database resource on the biology of disease from the Rosetta branch of the company. The genomic data will be managed by a new non-profit collaborative called Sage Bionetworks, which has Open Access as a core part of its mission. The emerging data and computational tools will be in the public domain for all to benefit as part of a growing open source biology movement.

A key aspect which should influence decisions on data sharing, participation and crowd-sourcing is quality: if a more open and creative, trade-based market will lead to better quality solutions, greater innovation and ultimately better science, then researchers could take advantage of these new business models. Each of us as citizens, are tax payers and indirectly contribute to funding science, so one could argue that greater participation in science developments by the general public, is a good thing and will lead to raised awareness, enhanced engagement and a more well-informed public; some of these notions are explored in a blog post on Open Research[73]. There are other benefits associated with capacity-building and workforce development, which may be accelerated by these participative approaches.

Consultation Challenge 3: Citizen Science

There are a number of basic questions in this area, which raise significant philosophical and pragmatic issues for professional scientists, research funding bodies, higher education institutions and the wider community, and some of them are presented here.

What are scientist and funder attitudes towards citizen science? What are the societal implications? What role should research funding bodies play?

What are the short, medium and long term strategic and policy implications on science practice and outcomes, of a more openly participative research approach which may pro-actively include the public?

What are the financial implications, both in terms of direct and indirect costs, investment in infrastructure and associated benefits? What are the risks? What is the impact on research quality (data, models, outcomes)?

Which disciplines and areas of research are most suited to citizen science methodologies? How should the collaboration market model be applied to research?

How will open and participative science initiatives impact on research practice in HE institutions? How should professional scientists, volunteers, amateurs and citizen scientists (and all flavours in between), work together in a socially optimal manner where there is mutual benefit? What can scientists learn from citizen journalism?

What are the technical requirements for designing effective citizen science Web services and systems? What can we learn from current successful exemplars?

8 Credentials, Incentives and Rewards

Credentials, reputation and recognition which act as incentives for scholarly research, are currently closely linked or dependant on the journal publishing model. Looking at this landscape more closely, we see that some very successful scientists are "proprietary", some scientists refuse to share their data[74], some scientists are blogging anonymously and there are success stories where scientists have achieved global recognition through open methodologies. Their working practices are featured in this Report.

8.1 Reputation and Trust

Reputation is multi-faceted but in the political economy of today's research assessment frameworks, the currency of choice is still the journal article, with citation data informing the assessment. The UK Research Excellence Framework (REF)[75] is at the second consultation stage and the final details of the REF will be issued during 2010, with the first REF exercise scheduled for 2013. The Consultation document states:

"We propose there be a maximum of either three or four outputs submitted for each researcher."

"All types of outputs from research that meets the Frascati principles (involving original investigation leading to new insights) will be eligible for submission. This includes 'grey literature' and outputs that are not in conventional published form, such as confidential reports to government or business, software, designs, performances and artefacts. Given that we see research as a process of investigation that has led to new insights effectively shared, we would expect all submitted work to include evidence of the research process, as well as presenting the insights in a form meeting the needs of its potential audience both within and beyond the academic community".

"We propose that citation information should be used in the REF as follows:

> *Citation data relating to submitted outputs will be provided to panels to inform expert review in UOAs covering the medical, health, biological and physical sciences, psychology, engineering and computer science. For other UOAs, panels should decide whether or not they would use citation information, after consulting their communities. We do not expect that the arts, humanities or many social sciences would opt to use citation information, given the limitations of such data in these subjects. "*

However, new notions of "reputation" are developing through metrics such as tracking the number of downloads, the number of citations to social Web entries, community ratings, recommender systems links, annotations and comments, the production of software and the generation and re-use of datasets.

In parallel, new notions of trust are emerging based on social network information, individual profiles and reviews. Cyber-Infrastructure Knowledge Networks On the Web (CI-KNOW)[76] is a suite of Web-based tools including a Recommender System, that facilitates the discovery of resources within communities. Recommendations arise from the analysis of social networks and structural linkages between items. These types of collective knowledge system combine the social and the semantic Web[77].

8.2 Incentivising Community Participation

The dependency of tenure, science career progression and advancement on attribution and credit together with the absence of an associated reward system linked to the social Web, is currently a powerful disincentive for many participative technologies and approaches, including

Open Notebook Science. The lack of incentives for bioinformatics annotation and curation has also been observed: assigning ratings to annotations (Harvana), measuring the volume of tags generated by individual postdocs and the use of an incentive points system with T-shirt prizes (NanoHub) have been tested, drawing on some of the successful reward strategies of online computer games.

The separate elements of the traditional science workflow have different relative values: the current journal publication process rewards original work published largely as formally-structured, text-based scholarly articles in journals. It does not explicitly reward data publication, the reuse of existing data, the subsequent analysis of that data or the application of models across that data. The growing volume of distributed datasets will require an associated expansion of computational data processing capability, and we may reasonably expect to see more data-driven discoveries with new knowledge arising wholly from data re-use, data analysis, large-scale simulations and complex modelling techniques. However, whilst the generation of data is increasing, there are indications that the level of data sharing and re-use is not.

A review of the RIN Report *To Share or Not to Share* (2008), suggests that "*a key policy imperative is to add to and reinforce the incentives and to reduce the constraints. Moreover, the Report suggests that there are risks in doing this in ways that do not recognise disciplinary differences*". [78] A new study of sharing and re-use of microarray data shows that: "*across 397 recent biomedical microarray studies, we found investigators were more likely to publicly share their raw dataset when their study was published in a high-impact journal, when their study was published in a journal with an enforceable data-sharing requirement, and when the first or last authors had high levels of career experience and impact"*.[79]

8.3 Measuring Contributions

We will also require new quality measures and indicators of success and impact for social Web contributions and open science activities: a new "**metrics of contribution**". How do you place a value on a blog post, a wiki entry, on open source code, on a dataset, on a participative project between professional scientists and citizens or on a YouTube video catalysing public engagement in science? Initiatives such as Metrics from Scholarly Usage of Resources (MESUR), have created a semantic model of scholarly communications based on the creation and analysis of a large scale semantic store of usage and citation information, leading to the formulation of guidelines and recommendations on impact metrics. PLoS ONE has published Article-level Metrics Information[80] which includes data from social bookmarks, blog coverage, Star ratings and usage data. In a more controversial step, a new **Scholar Factor (SF)**[81] has been proposed.

$$\begin{aligned}\text{SF} = {} & (\text{H Factor}) + \\ & (\text{Grant/Manuscript Review Factor}/20) + \\ & (\text{Annotations/Software/Datasets Factor}/5) \\ & + (\text{Web Factor}/50)\end{aligned}$$

The Scholar Factor includes an H Factor (as now but derived from Google Scholar data), a Grant/Manuscript Review Factor (based on data provided to grant funding agencies and journals), an Annotations/Software/Datasets Factor based on quantitative data for the number of authenticated annotations, software and gene sequences in open access archives, and a Web Factor, expressed as quantitative data covering authenticated blog posts, wiki postings etc. Clearly much more work on these types of microcredit-tracking/microattribution system(s)[82] is needed to fully capture the wealth and value of research contributions to the social Web and open data repositories.

Consultation Challenge 4: Credentials, Incentives and Rewards

There are many facets to discussion on open science and scholarly communications, incentivising data sharing and re-use, and on strategies for enabling more open participation.

Should open science practices be formally recognised and rewarded as intrinsic elements of scholarly communications? How can this be best achieved?

What are the views of the research community on appropriate incentives and reward structures for data sharing, data re-use and wider participation?

What are the views of the research funding bodies? Should these types of contribution and associated metrics, be included in future research assessment frameworks? How should they be assessed? How is the proposed Scholar Factor perceived? How should such metrics supplement journal citation metrics?

What are the views of scholarly publishers and learned societies? How do these contribution channels affect scholarly communication business models?

9 Institutional Readiness and Response

Team science implies (frequently distributed), partnerships of organisations, groups and individuals collaborating on large-scale data-driven research initiatives. In order to effectively support 21st Century team science, institutions need to have the appropriate research infrastructure in place.

9.1 Research Infrastructure

We can consider the primary elements of this infrastructure shown in Figure 3 to include:

- **Data** – We need to start to think about data as a utility for open sharing, recombination and re-use. Data and in particular reference datasets, should be viewed as vital infrastructure components and investments made to manage them accordingly. For the institution this means managing or curating the data generated by research staff, academics and faculty students.
- **Process** – These are the workflows, services, tools and methodologies that are used to capture, collect, process, combine, transform, mine, analyse, visualise, curate and preserve the data, models and simulations that are at the heart of large-scale data-driven science. Some of these processes will be overseen by the researchers themselves, but many services will be part of the portfolio of services managed and delivered by institutional information and computing services.
- **People** - Capacity and capability needs to be addressed at multiple levels: consortial, institutional, departmental/faculty, laboratory, group, research staff / post docs, postgraduate students, Library / Information Services, Computing Services etc. This human curation infrastructure is just as essential as the hardware, software and service components more usually associated with infrastructure definitions.

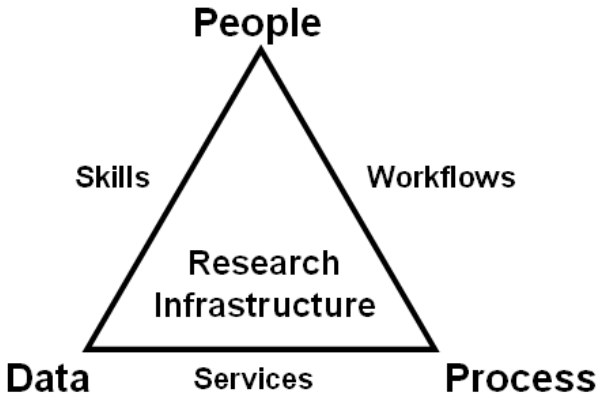

Figure 3: Research Infrastructure primary components

9.2 Organisational Structures, Planning and Policy

Many higher education institutions are reviewing their academic structures in the light of a range of environmental, political and economic drivers. To facilitate 21st Century open (team) science, institutions need to optimise their structures to enable inter-disciplinary and trans-disciplinary research, perhaps through the establishment of new inter-disciplinary research centres. Such centres may be physical, virtual or a distributed mix, and are where individuals and teams from international institutions and organisations in the public or private sectors, work collaboratively. These collaboratories, virtual research environments and globally distributed teams, require appropriate (social) networking infrastructure in the myExperiment mould, to support effective participant interaction, open sharing and discussion. As an illustration, the University of California, Los Angeles has created a new interdisciplinary centre focussing on high-throughput biology[83] to harness the power of technologies such as those from Pacific Biosciences, which enable high-throughput science. Other new centres resulting from recognition of the trend towards science at increasing orders of magnitude of scale, are likely.

This Report raises a number of fundamental issues for institutions associated with engaging with open science at Web scale. Firstly, there are basic questions around the strategic imperative for institutional senior management teams. Science at this scale must be on the radar of key research staff at PVC level and can be addressed through Research Committee agendas. Institutional Research Strategies and forward plans should address the key strategic, policy and operational issues in providing infrastructural support for these radical new data-driven research environments.

In addition, there are wider policy issues to consider. The University of Michigan has recently launched National Science Foundation funded Open Data Research Fellowships, however institutional support for the types of open science described in this Report has been queried during the fact-finding phase. Open science and data management policies need to be established which are informed by direction from the research funding councils and aligned to research assessment frameworks. Research work patterns may change: for example freelancing or the “Bursty Work” concept[84] may assist scientific collaborations, but this work approach may have implications for human resources departments.

There may be issues surrounding intellectual property rights (IPR) associated with research data, models and other derived outputs. In this context there may be tensions and potential conflict between university / institutional level legal requirements, licensing obligations and individual researcher’s philosophical aspirations.

A preliminary aide memoire for institutions is offered as an **Open Science Institutional Readiness Checklist** :

Open Science Institutional Readiness Checklist

1. *Open science principles addressed in Research Strategy?*
2. *Multi-scale research implications inform future planning?*
3. *Structures and processes to empower inter-disciplinary teams?*
4. *Position on professionals co-working with amateurs, volunteers and citizens?*
5. *Data sharing policy?*
6. *Research blogging/social networks policy?*
7. *Understanding of potential impact of new metrics of contribution?*
8. *LIS Director leading data advocacy programmes?*
9. *Faculty library staff providing data informatics support?*
10. *Data curation training embedded in research induction and DTC Programmes?*

Items 8-10 in the Checklist above, reference the role of Libraries and Information Services (LIS), and this is addressed in more depth in the next section.

Consultation Challenge 5: Institutional Readiness and Response

The open science agenda as well as the data-intensive science at extremes of scale described in this Report, have significant implications for higher education institutions at policy, planning and operational levels.

How aware are institutional senior management teams of the strategic implications of this potentially transformational agenda? How can research funding organisations, the JISC and other research support bodies help to raise awareness amongst institutional leaders? Who will lead and co-ordinate this work? What can be leveraged by partnerships on a global scale?

What are the implications for investment in research infrastructure? What can private sector organisations including ICT companies, contribute? What partnership opportunities arise?

How will academic structures evolve to support data-intensive science at extremes of scale? What institutional policy implications arise from open science practice? How are open scholarly communications channels such as research blogs supported in HEIs? Where are institutions positioned on open data-sharing? What are the IPR issues? What are the policy implications for institutions, of co-working with non-professionals i.e. volunteers and interested amateurs? What are the societal benefits?

What guidance is provided for research staff? How are open science issues and practices, addressed in staff induction and professional development courses? How can advocacy materials for institutions (e.g. a Team Science Toolkit), help to provide guidance and support for planning, policy development and good working practices?

10 Data Informatics Capacity and Capability

The skills mix required by the researcher of the future will include a substantive data informatics component and there are significant implications for curriculum development and modification at various levels, perhaps most importantly at postgraduate level. A report from the earlier eBank Project described how data manipulation and integration and associated skills, were required and used by students on the ChemInformatics course at the University of Southampton.[85] Whilst

the extent of this type of data informatics skills provision will vary with discipline, postgraduate student induction programmes, Doctoral Training Centres, and dedicated learning modules are possible delivery channels for training provision.

The importance of visualisations in data-intensive science was noted in Section 5.2: data visualisations can be used to good effect to convey complex information more clearly and there is intrinsic value in a compelling visualisation which can enhance reader/user understanding. Data visualisation skills are part of the skills set required within the science team, but what is the level of requirement, supply, demand and availability of data visualisation skills across different disciplines? Are visualisation skills taught within the new-entrant researcher curriculum? If such skills are in short supply, there are intriguing possibilities of extending the science team with graphic design experts, visual artists and others with graphics expertise from computer games, where visual impact is critical to success.

However, the new-entrant researcher is not the only role to require transformation. The Swan and Brown study[86] which arose from a Recommendation in the *Dealing with Data Report*, made a number of proposals associated with developing data science roles. An article[87] informed by presentations at the October 2008 ARL Forum on Re-inventing Science Librarianship noted that "*science librarians will need to become data consultants, data distributors, data service providers, data analysts, data miners and data curators*". Furthermore, the 2nd Research Data Management Forum[88] held in the UK in November 2008, also emphasised the pressing need for clarification on the roles and skills required by a "data workforce". [89] The Forum explored the variety of roles and responsibilities associated with effective data management. One outcome from the meeting was a synthesis of the core skills identified for the roles of data creator, data manager, data librarian and data scientist and these are summarised in Figure 4 below. Certain skills were shared across roles e.g. data preservation, metadata, data modelling, standards development.

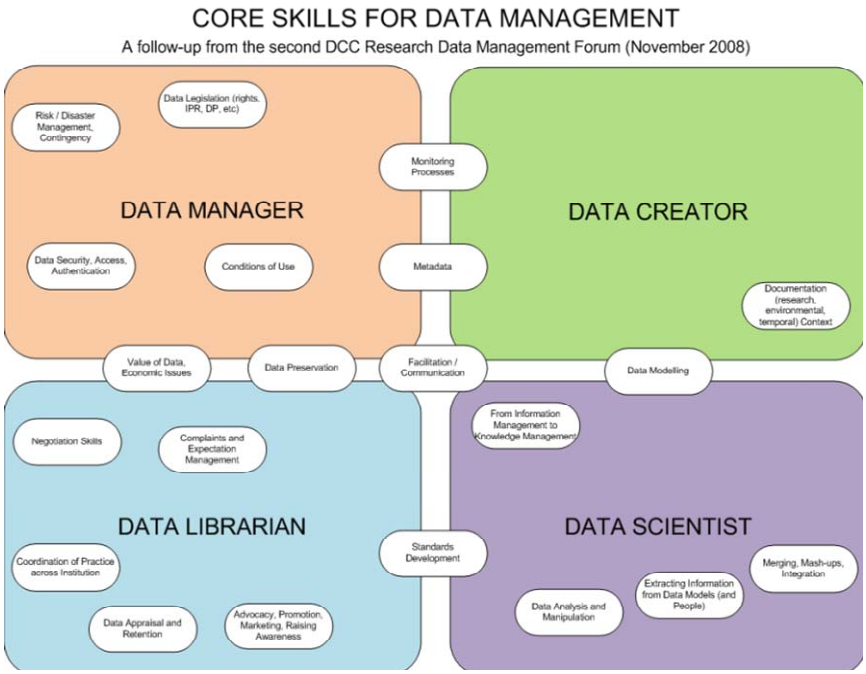

Figure 4 Core Skills for Data Management (Reproduced from IJDC)

21st Century open team science demands a fresh approach to the mix of skills and competencies required to do data-intensive / data-driven research. Expertise and skills from a range of disciplines come together in a new field of "**data Informatics**". A good exemplar of this skills mix underpinned the JISC-funded eBank and eCrystals Projects where domain scientists (chemists from the Chemistry Department at the University of Southampton), provided the core

subject knowledge. Computational science know-how came from the Department of Electronics and Computer Science at Southampton, whilst the informatics expertise was provided by UKOLN and the Digital Curation Centre at the University of Bath. Interestingly, some (but not all), of the informatics experts also had a bio-science background. Discussion in the early days of the eBank project was memorable in that literally hours were spent discussing and unpacking semantics in order to reach consensus, in particular certain words and concepts were interpreted differently even amongst this small research team: *data, metadata, data file* and *data-set* being the most troublesome.

10.1 Libraries and Research Data Management

A complementary analysis of data informatics functions based on the DCC Curation LifeCycle Model viewed from the Library perspective, was presented by the author in June 2009 at the ICSTI conference[90]. Ten key functions were identified which are shown in Figure 5 and summarised below.

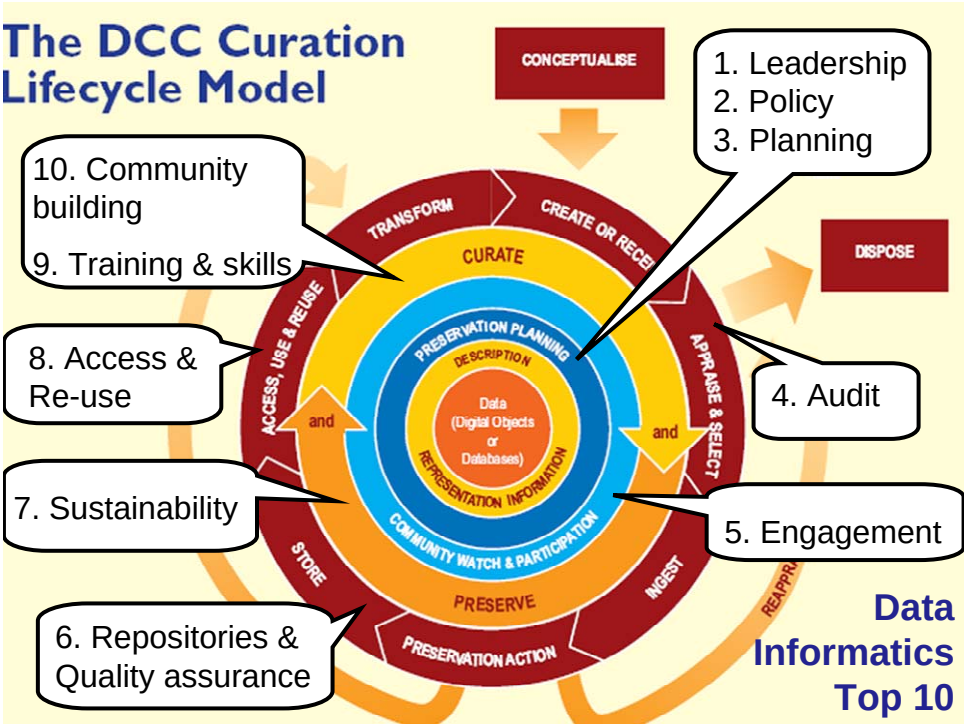

Figure 5 Data Informatics Top Ten for Libraries

1. **Leadership** : University Librarians and Directors of Information Services are very well-placed to demonstrate leadership for research data management within higher education institutions. Frequently they are in the institution Senior Management Team (SMT); they are able to advise the Vice-Chancellor or Principal of the strategic imperative for effective research data management; they may work closely with the PVC or VP Research on data policy development and they can liaise with IT Directors on the provision of data storage infrastructure. Furthermore, librarians are able to provide a co-ordination role for faculty data audits; they can raise awareness and carry out advocacy work through workshops, promoting best practice approaches and advising on curation lifecycle management. Finally they can deliver new support services to the local research community.
2. **Policy** : The DCC is monitoring UK and international policy development for research data management[91], however the JISC-funded Digital Preservation Policies Study by Neil Beagrie offered some high-level pointers and guidance for the development of local

institutional policies. An outline policy model / framework was derived with illustrative mappings to exemplars of other UK institutional strategies. Once again senior library staff are well-placed to draw on this work to inform and advise institutional policy makers.

3. **Planning** : Liaison and Faculty Librarians are positioned to advise faculty research staff, new-entrant researchers and postgraduates on the development of effective Data Management Plans (DMP). The DCC has developed a DMP Content Checklist[92] to assist with the process and is currently further developing this tool as part of an integrated community toolkit.

4. **Audit** : As part of the Data Audit Framework[93] development work, a number of pilot implementations were completed. Many of these were led by Information Services staff (e.g. University of Edinburgh) or Informatics staff (e.g. Centre for Computing in the Humanities at Kings College London, Innovative Design and Manufacturing Research Centre/UKOLN at the University of Bath). These staff can adopt valuable bridging roles and act as intermediaries and facilitators in executing the audit methodology.

5. **Engagement** : There is still a considerable risk that Libraries are perceived as passive observers offering remote support for open data-driven science rather than as integrated team players providing pro-active participation in the research process. There are a variety of models which can be followed ranging from simply extending the faculty / subject / liaison librarian role; seeking secondments to work as part of the research team *in situ* in the department; developing joint R&D projects with faculty members and carrying out immersive disciplinary "case studies" in order to gain a better understanding of the particular data issues. This latter approach was used by the DCC SCARP Project to good effect and a Report Synthesising the outcomes is forthcoming.

6. **Repositories** : There is a growing body of good practice guidance and advocacy materials around repository implementation. In many institutions, repositories have been implemented and are managed by library and information services staff. Whilst most repositories have to date focussed on collecting and storing textual documents, learning resources and multimedia materials, a growing number of institutions are beginning to tackle the challenge of data management. The DISC-UK Data Share initiative has published a useful Guide on Policy-making for Research Data in Repositories[94] and the Final Report[95] from the Project describes the experience of three institutions (the Universities of Edinburgh, Oxford and Southampton) seeking to enhance local data management practice.

7. **Sustainability** : There are various aspects to consider including the nature of a trusted repository and the ability to access and re-use datasets in the long-term. Faculty and research staff will want assurance of the robustness of the data infrastructure (local or remote) and there are a range of audit and certification frameworks (TRAC, DRAMBORA, NESTOR, Data Seal of Approval) for this purpose. For an exemplar, UKOLN/DCC as a partner in the eCrystals Project, has published three reports addressing aspects of the long-term preservation and sustainability of crystallography data sets stored in the eCrystals repository at Southampton. The papers cover Preservation Planning[96], Representation Information[97] and Preservation Metadata for Crystallographic Data[98]. They succinctly describe data preservation issues in one particular discipline based on the OAIS Reference Model components, including the RRORI Registry/Repository of representation information developed by the DCC in partnership with the EU-funded CASPAR Project and drawing on the PREMIS Data Dictionary.

8. **Access and Re-Use** : Community consensus on data formats, data standards, metadata schema and application profiles, use of domain identifiers etc. is critical to facilitating effective data access and re-use. Promoting the concept of Community Criteria for Interoperability from the earlier *Scaling Up Report*[99], is a useful first step which library staff can take in their advocacy work with faculty staff. In addition, advocating Linked Data principles for datasets[100] e.g. "RDF datasets with dereferenceable URIs".

9. **Training and Skills** : There are a growing number of courses and materials for developing digital curation and preservation expertise. The DCC has run a number of face-to-face DCC 101[101] Curation Training Courses based on the Curation LifeCycle Model and an online

version is in preparation. In addition, the Digital Preservation Training Programme[102] sponsored by the Digital Preservation Coalition and the Digital Curation Exchange[103] run by the University of North Carolina, provide teaching materials for professional development. There are real opportunities for libraries to cascade this knowledge to faculty staff and/or to initiate in-house mediated training to build local capacity and capability for data management.

10. **Community-building** : Library and information staff are very well-placed to facilitate local research community discussion around data management practice and associated challenges. More widely, the Research Data Management Forum cited earlier provides an opportunity for diverse stakeholders to meet and examine selected data themes whilst there are various conferences (including the well-established International Digital Curation Conference[104] now in its 5th year), for networking and exchange of experience.

10.2 New Roles, New Skills, New Curricula

Whilst more survey and analysis work would be required to map the frequency, level, and scope of informatics training opportunities for postgraduates within disciplinary curricula more widely, there is some evidence[105] that (in the UK at least), there is an urgent need to build capacity and capability in data informatics across a range of roles (data librarians, data scientists, data creators etc.). Whilst these types of skills are essential embedded elements within selected disciplinary curricula (e.g. bioinformatics, chem-informatics, health informatics etc.), there are relatively few data-oriented elements to library and information science curricula in the UK. This is one opportunity when data informatics skills can be introduced and developed in a structured fashion and a good example of such a course is the ALA-accredited Masters course at the Graduate School of Library and Information Science at the University of Illinois, which has a specialisation in Data Curation[106].

The thorny issue of career progression for individuals with informatics expertise will be greatly advanced by achieving a critical mass of research and support staff in higher education with these essential skills. Given the earlier discussion about new metrics and incentives, data informatics contributions to the body of scholarly information, should receive the recognition they deserve. As a useful step in this direction, a new international society for biocuration was launched at the 3rd International Biocurators Conference in April 2009. The mission of the Society is reproduced below:

1. *Define the work of biocurators for the scientific community and the public funding agencies;*
   2. Propose a discussion forum for interested biocurators, developers, scientists and students.
   3. Organize a regular meeting where biocurators will be able to present their work and discuss their projects.
   4. Lobby to obtain increased and stable funding for biocuration resources that are essential to research;
   5. Build a relationship with publishers and establish a link between researchers and databases through journal publishers
   6. Organize a regular workshop where new biocurators, or interested students can be trained in the use of the common tools needed for their work.
   7. Provide documentation on the use of common database and bioinformatics tools.
   8. Provide 'Gold Standards' for databases, such as the use of unique, traceable identifiers, use of shared tools, etc.;
   9. Share documentation on standards and annotation procedures with the aim of developing Standard Operating Procedures (SOPs).
   10. Foster connections with user communities to ensure that databases and accompanying tools meet specific user needs;
   11. Maintain a biocurator job market forum.

There are clear connections between biocurator roles and the proposed transformational role of librarians, and indeed there is a "Biolibrarian" proposal[107] put forward by scientists from the Biotechnology Centre, Oslo.

The role of the library and information service is critical. Libraries can lead the provision of advocacy services for data curation to support the research agenda. They should be fully-integrated partners in the team. Libraries can provide guidance in all the ways outlined above (see University of Edinburgh example[108]). Information services staff can help to shape policy for consortial teams who wish to employ open science methods; they can advise on the choice of tools and platforms; they can provide technical expertise on the embedding of standards, metadata, schema, data models and terminologies. They can also contribute to the induction and training of research staff. Data curation modules should be embedded in the new-entrant programmes run by Doctoral Training Centres (DTC), making use of the rich materials and resources produced by the Digital Curation Centre. Libraries need to act now, engage and pro-actively participate in open data-driven team science.

Consultation Challenge 6: Data Informatics Capacity and Capability

There are a number of issues associated with the embedding of skills required for open data-intensive science and the role of the Library and Information Services. There are also implications for postgraduate training and LIS curriculum development.

What is the research community view on the current provision of data informatics skills for postgraduates and research staff? If current curricula and training are not meeting needs, how can the position be improved? Should basic data informatics training be a core element of courses? Who should provide this training? What are the costs?

How can research funding agencies best support data informatics skills development?

What is the community perspective on the roles that Libraries and Information Services could play in supporting open data-intensive science? How can academic and research libraries be empowered to engage and participate in team science initiatives?

What is the role of SCONUL, RLUK, CILIP and other professional LIS organisations?

How should Library and Information Science schools address the provision of data curation and data informatics expertise within their courses and programmes?

11 Conclusions

This Report has attempted to draw together and synthesise evidence and opinion from a wide range of sources. Examples of data intensive science at extremes of scale and complexity which enable forecasting and predictive assertions, have been described together with compelling exemplars where an open and participative culture is transforming science practice. It is perhaps worth noting that the pace of change in this area is such, that it has been a challenging piece to compose and at best, it can only serve as a subjective snapshot of a very dynamic data space. However, the Report has raised many questions and challenges associated with open science, for a wide range of stakeholder organisations. It is hoped that the presentation of the Report as a Consultative document, will be instrumental in stimulating further discussion and debate both within and beyond the sector.

The perspective of openness as a continuum is helpful in positioning the range of behaviours and practices observed in different disciplines and contexts. By separating the twin aspects of openness (access and participation), we can begin to understand the full scope and potential of the open science vision. Whilst a listing of the perceived values and benefits of open science is given, further work is required to provide substantive and tangible evidence to justify and

support these assertions. Available evidence suggests that transparent data sharing and data re-use are far from commonplace. The peer production approaches to data curation which have been described, are really in their infancy but offer considerable promise as scaleable models which could be migrated to other disciplines. The more radical open notebook science methodologies are currently on the “fringe” and it is not clear whether uptake and adoption will grow in other disciplines and contexts.

Whilst there are established exemplars of effective citizen science, this model may be more suited to certain domains and types of research. However, the growth of mobile phone use in citizen journalism and the continuously enriched functionality of mobile devices, suggest that there is great potential for more participatory methodologies to benefit scientific research, though some privacy and legislative issues remain unanswered.

The influence of computer gaming approaches on volunteer computing initiatives to motivate participants is noteworthy, and there is scope for wider adoption of such tactics. The development of citizen science Web services, system architectures and the design of appropriate interfaces, is still at a relatively early stage. We need to learn much more about how the public interact with these services to maximise the value and benefit from such investment.

The potential impact of these changing practices on established business models for science and scholarly communications has been identified. It has already been noted that data sharing and re-use is relatively limited, however new notions of reputation and trust are developing which challenge established norms. The current journal publishing model with associated citation metrics for research assessment, does not reward data sharing, social Web contributions or peer production approaches to data curation. Against this background, novel proposals are appearing which seek to include such parameters in research assessment metrics, but the implications on research funder policies, future science investment planning and scholarly communication business models are not fully understood. It is clear however, that the lack of incentives for data sharing and participatory methodologies, are a barrier to the wider adoption of the open science agenda.

The implications of open science practice on higher education institutions are many and varied, and this Report has done no more than raise some preliminary points. However it is hoped that by asking basic questions which explore institutional awareness, policy, planning and research practice, the community will begin to explore these substantive issues in more depth. Particular attention has been paid to the provision of data informatics capacity and capability and the role of the Library in this context. The Report asserts that Libraries are well-placed to support research data management but that new skills and roles will need to be embraced by the professional LIS community. Modifications to LIS courses will be required and there are similar training implications for new-entrant researchers and postgraduates, to equip them with the skills and methodologies required for data-intensive science. The UK Digital Curation Centre is a key resource, although the increasing demands on this relatively modest service are challenging.

Finally, it is hoped that this Report will stimulate and contribute to community discussion in the UK, but also fuel the open science debate on the global stage. The potential impact of data-intensive open science on research practice and research outcomes is both substantive and far-reaching. The issues raised here will require fuller articulation and investigation. The economic implications will require detailed analysis and the societal benefits should be reviewed and evaluated. However for now, it is sufficient to draw together a number of associated themes (possibly for the first time), to stimulate and foster wider debate and to challenge the community to fully explore the transformational potential of open science at Web-scale.

12 Appendix: Contributors

Interviews, discussions and presentations from the following individuals contributed to the data collection phase of this consultancy:

Dan Atkins, University of Michigan

Fran Berman, RPI

Jean-Claude Bradley, Drexel University

Simon Coles, University of Southampton

Nosh Contractor, NorthWestern University

Stephen Emmott, Microsoft Research

Jeremy Frey, University of Southampton

Carole Goble, University of Manchester

Chris Greer, NITRD

Timo Hannay, Nature Publishing Group

Tony Hey, Microsoft Research

Jane Hunter, University of Queensland

Clifford Lynch, CNI

Cameron Neylon, STFC

Andrew Treloar, ANDS

John Wilbanks, Science Commons

13 References

[1] Frank Gibson, Do scientists really believe in open science? http://peanutbutter.wordpress.com/2007/06/26/do-scientists-really-believe-in-open-science/

[2] Bill Hooker, The Future of Science is Open
http://3quarksdaily.blogs.com/3quarksdaily/2007/01/the_future_of_s.html

[3] Science Commons Principles for Open Science 2008
http://sciencecommons.org/resources/readingroom/principles-for-open-science/

[4] Amazon S3 http://aws.amazon.com/s3/

[5] Lorcan Dempsey's blog. Web Scale http://orweblog.oclc.org/archives/001238.html

[6] J-C Bradley http://drexel-coas-elearning.blogspot.com/2006/09/open-notebook-science.html

[7] Predictive Science Academic Alliance Program (PSAAP) http://www.sandia.gov/NNSA/ASC/univ/psaap.html

[8] Carole Lartigue et al, Creating Bacterial strains from genomes that have been cloned and engineered in yeast. Science (2009) http://www.sciencemag.org/cgi/content/full/325/5948/1693

[9] Justin Rattner. Immersive Science. http://bx.businessweek.com/intel-vs-amd/view?url=http%3A%2F%2Fblgint.infosyssaas.com%2Fresearch%2F2008%2F11%2Fimmersive_science.php

[10] Wuchty S., Jones, BF, Uzzi B. (2007) The Increasing Dominance of Teams in Production of Knowledge. Science Vol 316 (5827), pp 1036-1039. http://www.sciencemag.org/cgi/content/abstract/1136099

[11] Tony Hey and Anne Trefethen. The Data Deluge: An eScience Perspective (2003) http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf

[12] Microsoft Research 2006. Towards 2020 Science. http://research.microsoft.com/towards2020science/background_overview.htm

[13] Anderson, C. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired Magazine* July 16th, 2008. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

[14] Nature Special Issue 4th September 2008 http://www.nature.com/news/specials/bigdata/index.html

[15] Cluster Exploratory Programme at the National Science Foundation. http://www.nsf.gov/pubs/2008/nsf08560/nsf08560.htm

[16] Cloudera http://www.cloudera.com/

[17] RIN Stewardship Principles (2007) http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines

[18] Liz Lyon (2007) Dealing with Data Report. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

[19] UKRDS Final Report (2009) http://ukrds.ac.uk/resources/

[20] Cacioppo J. The Rise in Collaborative Psychological Science. APS Observer, Vol 20 (9), 2007. http://www.kellogg.northwestern.edu/faculty/uzzi/ftp/Observer%20column%20-%20October2007.pdf

[21] Fiore, SM, (2008). Interdisciplinarity as Teamwork: How the Science of Teams can inform Team Science. Small Group Research 39, 251-277. http://sgr.sagepub.com/cgi/content/refs/39/3/251

[22] John Whitfield. Group theory (2008) Nature. http://www.nature.com/news/2008/081008/full/455720a.html

[23] Virtual Research Environments. http://www.jisc.ac.uk/whatwedo/programmes/vre.aspx

[24] Shirley Wu. New job and curation 101 http://shirleywho.wordpress.com/2009/09/14/new-job-and-curation-101/

[25] Art Rai and James Boyle. Synthetic biology: caught between property rights, the public domain and the commons. PLoS Biology (2007) http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0050058

[26] Bryn Nelson. Data sharing: Empty Archives. Nature (2009) http://www.nature.com/news/2009/090909/full/461160a.html

[27] The New Science of Metagenomics: Revealing the Secrets of our Microbial Planet. National Academies Press http://www.nap.edu/catalog.php?record_id=11902#toc

[28] Gene Machine http://www.pacificbiosciences.com/assets/files/Forbes_GeneMachine091809.pdf

[29] Infrastructure for Integration in Structural Sciences (I2S2) http://www.ukoln.ac.uk/projects/I2S2/

[30] Core Scientific Metadata Schema (CSMD) http://epubs.cclrc.ac.uk/bitstream/485/

[31] Curation LifeCycle Model http://www.dcc.ac.uk/lifecycle-model/

[32] Y. Setty et al, Four-dimensional realistic modelling of pancreatic organogenesis. PNAS (2008) http://www.pnas.org/content/105/51/20374.abstract

[33] Douglas Kell, The virtual human: towards a global systems biology of multiscale, distributed biochemical network models. (2007) http://www3.interscience.wiley.com/journal/117888899/abstract

[34] Toby Segaran & Jeff Hammerbacher. "Beautiful Data". Publ. O'Reilly (2009) .

[35] Stuart Macdonald. Web 2.0 Data Visualisation Tools Part 1Numeric Data http://ie-repository.jisc.ac.uk/226/1/Numeric_data_mashup.pdf and Part 2 Spatial Data http://ie-repository.jisc.ac.uk/305/1/spatial_data_mashup_V2.pdf

[36] Tara Matthews et al, Designing Glanceable peripheral displays. http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-113.pdf

[37] Angus Whyte, Curating Brian Images in a Psychiatric Research Group, DCC SCARP Project Report http://www.dcc.ac.uk/docs/publications/case-studies/SCARP_B4821_NeuroCase_v1_1.pdf

[38] Franck Capello et al. Toward exascale resilience. (2009) http://jointlab.ncsa.illinois.edu/pubs/Toward_Exascale_Resilience.pdf

[39] GODIVA2 visualisation demo page http://behemoth.nerc-essc.ac.uk/ncWMS/godiva2.html

[40] Minimal Information Requested in the Annotation of Models (MIRIAM) http://www.ebi.ac.uk/miriam/main/mdb?section=standard

[41] Jeremy Ginsberg et al, Detecting influenza epidemics using search engine query data. (2009) http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html

[42] Michael Nielsen, The Future of Science http://michaelnielsen.org/blog/?p=448

[43] http://www.wired.com/science/discoveries/magazine/15-10/st_essay

[44] Heather Piwowar. Measuring the adoption of open science. http://www.slideshare.net/hpiwowar/measuring-the-adoption-of-open-science-presentation

[45] A critical analysis of social networking sites for scientists http://docs.google.com/View?docid=dhs5x5kr_572hccgvcct

[46] Google Wave demo http://blog.openwetware.org/scienceintheopen/2009/08/23/reflecting-on-a-wave-the-demo-at-science-online-london-2009/

[47] Waldrop, M.M. Science 2.0 – Is Open Access Science the Future? http://www.sciam.com/article.cfm?id=science-2-point-0

[48] Science Online London 2009 http://www.scienceonlinelondon.org/index.php

[49] The Open Laboratory http://scienceblogs.com/clock/2009/03/the_open_laboratory_2008_is_he.php

[50] Howe D. et al Big data: the future of biocuration (2008) Nature http://www.nature.com/nature/journal/v455/n7209/full/455047a.html

[51] Barend Mons et al. Calling on a million minds for community annotation in WikiProteins. Genome Biology (2008) http://genomebiology.com/2008/9/5/R89

[52] Pico A.R. et al WikiPathways: Pathway editing for the people (2008) PLoS Biology http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0060184

[53] Concept Web Alliance http://conceptweballiance.org/

[54] The Curators Manual for ChemSpider http://www.chemspider.com/docs/The_Curators_Manual_for_ChemSpider.pdf

[55] Cory Doctorow. Interviewed by Richard Poynder. http://www.earlham.edu/~peters/fos/2006/04/richard-poynder-interviews-cory.html

[56] Murray-Rust, P. Chemistry for everyone. http://www.nature.com/nature/journal/v451/n7179/full/451648a.html

[57] UsefulChem http://usefulchem.wikispaces.com/

[58] Bradley, J-C et al Nature Precedings (2008) http://precedings.nature.com/documents/2237/version/1

[59] ChemTools http://blogs.chem.soton.ac.uk/

[60] Cameron Neylon, Head in the Clouds – Automated Experimentation (2009) http://cameronneylon.wikidot.com/head-in-the-clouds-automated-experimentation

[61] J-C Bradley et al. (2009) Beautifying data in the real world. Chapter in Beautiful Data, Publ O'Reilly.

[62] BBC LabUK http://www.bbc.co.uk/labuk/

[63] eBird Ithaka Case Study http://www.ithaka.org/ithaka-s-r/strategy/ithaka-case-studies-in-sustainability/case-studies/SCA_BMS_CaseStudy_eBird.pdf

[64] Dan Schultz How citizen journalists can learn from the work of "citizen scientists" http://www.pbs.org/idealab/2009/08/how-citizen-journalists-can-learn-from-work-of-citizen-scientists238.html

[65] Eric Paulos et al Citizen Science: Enabling participatory urbanism (Book chapter in Press) http://www.eecs.berkeley.edu/~honicky/CitizenScience.pdf

[66] David M. Aanensen et al. EpiCollect: Linking Smartphones to Web applications for Epidemiology, Ecology and Community Data Collection. PLoS ONE (2009) http://www.plosone.org/article/info:doi/10.1371/journal.pone.0006968

[67] 12 Worthy Causes Seek Your Spare PC Cycles. PC World (2009) http://www.pcworld.com/printable/article/id,171126/printable.html

[68] Foldit game http://fold.it/portal/

[69] Jane Hunter et al HarVANA – Harvesting Community Tags to enrich collection metadata , JCDL (2008) http://portal.acm.org/citation.cfm?id=1378889.1378916

[70] Von Ahn L. et al reCAPTCHA: Human-based character recofnition via Web security measures Science (2008) http://www.sciencemag.org/cgi/content/full/321/5895/1465

[71] Open Notebook Science Challenge http://onschallenge.wikispaces.com/

[72] Michael Neilsen http://michaelnielsen.org/blog/the-future-of-science-2/

[73] Cameron Neylon. Open Research: the personal, the social and the political. http://blog.openwetware.org/scienceintheopen/2009/10/10/open-research-the-personal-the-social-and-the-political/

[74] Caroline Savage and Andrew Vickers. Empirical study of data sharing by authors publishing in PLoS journals. (2009) http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0007078

[75] HEFCE Research Excellence Framework http://www.hefce.ac.uk/Research/ref/

[76] Huang, Contractor and Yao CI-KNOW: Recommendation based on Social Networks, Proc Int Conf Digital Government Research (2008) http://portal.acm.org/citation.cfm?id=1367840

[77] Tom Gruber, Collective Knowledge Systems: Where the Social Web meets the Semantic Web (2007) http://tomgruber.org/writing/CollectiveKnowledgeSystems.pdf

[78] Aaron Griffiths, The Publication of Research Data: Researcher Attitudes and Behaviour (2009). IJDC http://www.ijdc.net/index.php/ijdc/article/view/101

[79] Heather Piwowar. Public sharing of research datasets: a pilot study of associations. In:ASIS&T and ISSI Pre-Conference: Symposium on Informetrics and Scientometrics (2009). Abstract: http://www.researchremix.org/wordpress/publications/

[80] Mark Patterson. Measuring Impact where it matters, PLoS ONE (2009) http://www.plos.org/cms/node/478

[81] Bourne and Fink, I am not a Scientists, I am a Number, PLoS Computational Biology (2009) http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000247

[82] Incentives/rewards for scientific contribution http://www.gen2phen.org/researcher-identification-primer/incentivesrewards-scientific-contributions

[83] UCLA Center for High Throughput Biology http://chtb.bioinformatics.ucla.edu/

[84] Deepak Singh. The future of scientific collaboration: extending the “Bursty Work” concept. http://mndoci.com/2008/02/03/the-future-of-scientific-collaboration-extending-the-bursty-work-concept/

[85] Grainne Conole. External evaluation of the eBank Project Report (2006). http://www.ukoln.ac.uk/projects/ebank-uk/evaluation-report-dec-2006/evaluation-report-december-2006.pdf

[86] Alma Swan and Sheridan Brown (2008) Skills, role and career structure of data scientists and curators : Assessment of current practice and future needs Report. http://www.jisc.ac.uk/publications/documents/dataskillscareersfinalreport.aspx

[87] ARL Research Library Issues February 2009, No 262, Elisabeth Jones, Reinventing Science Librarianship http://www.arl.org/bm~doc/rli-262-science.pdf

[88] 2nd Research Data Management Forum http://www.dcc.ac.uk/events/data-forum-2008-november/

[89] Graham Pryor and Martin Donnelly. Skilling up to do data: Whose role? Whose responsibility? Whose career? .IJDC 4(2), (2009) (in press).

[90] Liz Lyon Libraries and team Science Presentation at ICSTI Conference 2009, Ottawa, Canada http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html#ottawa-jun-2009

[91] UK Curation Policies http://www.dcc.ac.uk/resource/curation-policies/

[92] DCC Data Management Plan Content Checklist http://www.dcc.ac.uk/docs/templates/DMP_checklist.pdf

[93] Data Audit Framework http://www.data-audit.eu/users.html

[94] DISC-UK Data Share Guide to Policy Making for research Data in Repositories http://www.disc-uk.org/docs/guide.pdf

[95] DISC-UK DataShare Project Final Report http://ie-repository.jisc.ac.uk/336/1/DataSharefinalreport.pdf

[96] Manjula Patel, Preservation Planning for Crystallography Data UKOLN/DCC 2009 http://wiki.ecrystals.chem.soton.ac.uk/images/8/82/ECrystals-WP4-PP-090625.pdf

[97] Manjula Patel, Representation Information for Crystallography Data UKOLN/DCC 2009 http://wiki.ecrystals.chem.soton.ac.uk/images/8/82/ECrystals-WP4-PP-090625.pdf

[98] Manjula Patel, Preservation Metadata for Crystallography Data, UKOLN/DCC 2009 http://wiki.ecrystals.chem.soton.ac.uk/images/9/9d/ECrystals-WP4-PM-Final.pdf

[99] Liz Lyon, Simon Coles, Monica Duke, Traugott Koch. Scaling Up Report (2008) http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/Ebank3report.pdf

[100] LinkedData for datasets http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets

[101] DCC Curation 101 Training Course http://www.dcc.ac.uk/events/digital-curation-101-2008/

[102] Digital Preservation Training Programme http://www.dptp.org/

[103] Digital Curation Exchange http://digitalcurationexchange.org/

[104] International Digital Curation Conference http://www.dcc.ac.uk/events/dcc-2009/

[105] Stuart Macdonald and Luis Martinez-Uribe, Data librarianship – a gap in the market, CILIP 2008 http://www.cilip.org.uk/publications/updatemagazine/archive/archive2008/june/Interview+with+Macdonald+and+Martinez-Uribe.htm

[106] Master of Science, Specialisation in Data Curation http://www.lis.illinois.edu/programs/ms/data_curation.html

[107] Biolibrarian proposal, Biotechnology Centre, Oslo http://irefindex.uio.no/wiki/The_Biolibrarian_Proposal

[108] University of Edinburgh. Research data management guidance. http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt